НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»
Інститут прикладного системного аналізу
Кафедра математичних методів системного аналізу

«На правах рукопису»
УДК 004.855.5

«До захисту допущено»
Завідувач кафедри
_____ О.Л.Тимощук
«___»_____2018 р.

Магістерська дисертація

на здобуття ступеня магістра
зі спеціальності 124 Системний аналіз
на тему: Інтелектуальна система розпізнавання тональності тексту

Виконав:
студент ІІ курсу, групи КА-61м
Шипік Данило Володимирович _____
Керівник:
к.т.н., доц., доцент кафедри ММСА
Дідковська М.В. _____
Рецензент:
доцент кафедри програмного
забезпечення комп'ютерних
систем факультету прикладної
математики КПІ імені Ігоря Сікорського,
к.т.н., доц. Заболотня Т.М. _____

Засвідчую, що у цій магістерській дисертації немає запозичень з праць інших авторів без відповідних посилань.
Студент _____

Київ
2018

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Інститут прикладного системного аналізу
Кафедра математичних методів системного аналізу

Рівень вищої освіти — другий (магістерський)
Спеціальність (спеціалізація) — 124 «Системний аналіз» («Системний аналіз і управління»)

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ О.Л.Тимощук
«___»_____20__ р.

ЗАВДАННЯ
на магістерську дисертацію студенту

Шипік Данило Володимирович

1. Тема дисертації: Інтелектуальна система розпізнавання тональності тексту, науковий керівник дисертації Дідковська Марина Віталіївна, к.т.н., доц., затверджені наказом по університету від 27.03.2018 р. № 1028-с

2. Термін подання студентом дисертації: 18.05.2018 р.

3. Об’єкт дослідження: Обробка природної мови у контексті методів розпізнавання тональності тексту

4. Предмет дослідження: Методи розпізнавання тональності коментарів до новинних статей, як коротких, емоційно забарвлених текстів зі змінюваною лексикою

5. Перелік завдань, які потрібно розробити: Проаналізувати існуючі методи та рішення для визначення тональності текстів. Вивчити можливості збільшення ефективності існуючих методів. Розробити програмний продукт, що вілює запропонований метод. Проаналізувати ефективність запропонованих покращень. Розробити стартап-проект на основі створеного програмного забезпечення.

6. Орієнтовний перелік графічного (ілюстративного) матеріалу: Існуючі методи аналізу тональності текстів та аналіз їх переваг та недоліків; Модифікований метод аналізу тональності текстів; Практичний приклад роботи програми.

7. Орієнтовний перелік публікацій: стаття у міжнародному науковому журналі: Порівняння точності алгоритмів аналізу тональності на прикладі твіттів

8. Дата видачі завдання: 16.03.2018 р.

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Термін виконання етапів магістерської дисертації	Примітка
1.	Підготовка та оформлення вступу	18.03.2018	
2.	Підготовка і оформлення першого і другого розділів	28.03.2018	
3.	Підготовка і оформлення третього розділу	10.04.2018	
4.	Підготовка і оформлення четвертого розділу	20.04.2018	
5.	Підготовка і оформлення концептуальних висновків	04.05.2018	
6.	Підготовка і оформлення презентації доповіді на захисті	14.05.2018	

Студент Д.В.Шипік

Науковий керівник дисертації М.В.Дідковська

РЕФЕРАТ

Магістерська дисертація: 75 с., 14 рис., 26 табл., 2 додатки, 4 розділи та 17 джерел.

Об’єктом дослідження є методи розпізнавання тональності тексту.

Метою даної роботи є розробка модуля для оцінювання ставлення до новин за тональністю коментарів, а також реалізація додатку для демонстрації роботи модуля. У роботі проаналізовано методи аналізу тональності, проведено огляд існуючих підходів.

Результати роботи:

- запропоновано модифікацію алгоритму аналізу тональності для коментарів, як коротких, емоційно забарвлених текстів зі змінюваною лексикою;
- реалізовано запропоновану модифікацію класифікатора;
- створено чат бота із зручним для користувача інтерфейсом.

Новизна роботи:

- запропоновано принципово новий спосіб використання аналізу тональності, а саме аналіз емоційної реакції на певні новини за їх коментарями та подальше надсилання новин за допомогою чат-бота;
- обґрунтовано використання власних підходів до розв’язку задачі аналізу тональності.

Результати даної роботи рекомендується використовувати для аналізу емоційної реакції користувачів на певні новини, теми, визначення ставлення аудиторії певних сайтів до певних новин, персоналізації новинного сектору

ОБРОБКА ПРИРОДНОЇ МОВИ, АНАЛІЗ ТОНАЛЬНОСТІ ТЕКСТУ, НАЇВНИЙ БАЙЄСОВСЬКИЙ КЛАСИФІКАТОР, МЕТОД ОПОРНИХ ВЕКТОРІВ, НАПІВАВТОМАТИЧНЕ НАВЧАННЯ.

ABSTRACT

Master thesis: 75 p., 14 fig., 26 tabl., 2 appendixes, 4 sections, N sources.

The aim of this work is the development of a software module designed to evaluate the emotional response to news using comments, as implementation of the application to demonstrate the capabilities of the module. The thesis analyzes concepts, approaches and methods of sentiment analysis.

The results of the thesis:

- the modification of algorithm for analysis of the sentiment of the news comments as short, emotionally colored texts with changing vocabulary is proposed;
- the proposed modification of the classifier is implemented;
- a chat bot with a user-friendly interface is created

Novelty of work:

- principally new approach of using sentiment analysis is proposed: analysis of emotional response to the news by their comments and further sending of news via chat-bot;
- the usage of own approaches to the decision of problems of compression of images is substantiated.

Results of this work can be used for assessment users' emotional reaction on the news, topics, evaluation of certain sites auditory opinion on the news, determining the attitude of the audience of certain sites to certain news, personalization of the news sector.

NATIVE LANGUAGE PROCESSING, SENTIMENT ANALYSIS, NAÏVE BAYES CLASSIFIER, SVM, SEMI-SUPERVISED LEARNING.

ЗМІСТ

ВСТУП

Думки інших людей завжди були важливою частиною інформації для більшості із нас в процесі прийняття рішень. Задовго до того, як використання Всесвітньої мережі стало невід'ємною частиною повсякденного життя суспільства, люди запитували поради в інших щодо якості тієї чи іншої техніки, послуг тієї чи іншої фірми або якому претенденту їх знайомі віддають перевагу на місцевих виборах. Широке розповсюдження Інтернету зробило можливим ознайомлення з поглядами та досвідом великої кількості людей по всьому світу он–лайн, і, навпаки, все більше і більше людей роблять свої думки та погляди доступними для широкої громади.

Інтерес, який демонструють користувачі до он–лайн відгуків та коментарів, а також потенціальний вплив цих коментарів на питання у дискурсі і прийняття рішень змушують звернути увагу на цей аспект он–лайн активності. З появою компонентів Web 2.0, таких як блоги, форуми, соціальні мережі та багато інших типів соціальних медіа споживачі мають у своєму розпорядженні плацдарм для висловлення своїх думок, позитивних та негативних щодо будь–якої новини чи ідеї. З початку 2000 року, аналіз тональності зріс до однієї з найбільш активно досліджуваних галузей в області обробки природної мови, з'явились великі корпоративні проекти, пов'язані з аналізом тональності і також величезна кількість різноманітних стартапів. Багато великих корпорацій створили свої власні внутрішні рішення для проведення аналізу настроїв.[1]

Дослідження в області аналізу тональності знаходяться на початковій стадії, незважаючи на зростаючі потреби суспільства в аналізі соціальних думок. Окрім того, існує досить багато проблем, з якими стикаються дослідники при розробці методів автоматичного аналізу тональності.

Не зважаючи на численні способи використання аналізу тональності, проектів, які б зосередили свою увагу на емоційному відклику на новини надзвичайно мало.

Більшість дослідників та розробників обрали об’єктом своєї роботи аналіз техніки, фільмів, ресторанів та іншого.

Однак аналіз новинних сайтів може стати дуже цінним ресурсом для політ-технологів, маркетологів та працівників мас медіа, для яких реакція на новини є надзвичайно важливою.

Таким чином, метою дослідження є розробка системи оцінювання емоційного відклику на новини за коментарями, опис її принципів роботи та її практична реалізація.

Для досягнення такої мети були вирішені наступні задачі:

- проаналізувати існуючі методи та рішення у галузі визначення тональності текстів;
- вивчити можливість збільшення ефективності існуючих методів;
- розробити програмний продукт, що втілює запропонований метод;
- проаналізувати ефективність запропонованих покращень;
- розробити стартап-проект на основі створеного ПЗ.

Об’єктом дослідження є методи розпізнавання тональності тексту.

Предметом дослідження є методи Наївного Байеса, метод опорних векторів та метод Хе Юлан та Жоу Деу у контексті розпізнавання тональності текстів.

Наукова новизна роботи полягає у наступному:

- запропоновано принципово новий спосіб використання аналізу тональності, а саме аналіз емоційної реакції на певні новини за їх коментарями та подальше надсилання новин за допомогою чат-бота;
- обґрунтовано використання власних підходів до розв’язку задачі аналізу тональності.

Апробація роботи. За результатами даної роботи була підготовлена стаття «Порівняння точності алгоритмів аналізу тональності на прикладі твіттів», яка була опублікована у журналі «Інтернаука».

Було запропоновано алгоритм аналізу тональності, що базується на методах Наївного байєса та напівавтоматичного навчання.

Практичними результатами роботи є реалізація телеграм-бота, що надсилає підбірку позитивних з точки зору користувача новин, використовуючи розроблений модуль розпізнавання тональності коментарів.

Робота складається з 4 розділів. В першому розділі розглядається актуальність проблеми та існуючі підходи до її розв’язання, формалізується постановка задачі дослідження. Другий розділ присвячений математичним основам методів, які було застосовано у дослідженні, вибору критеріїв оцінки результатів роботи та модифікації існуючих методів. У третьому розділі здійснено огляд технологій та аргументовано вибір інструментів, що використовуються в роботі, описано процес розробки програмного продукту та проведено дослідження отриманих результатів його роботи. Четвертий розділ присвячено розробленому на основі даної магістерскої дисертації стартап-проекту.

1 ОГЛЯД ЗАДАЧІ АНАЛІЗУ ТОНАЛЬНОСТІ

1.1 Аналіз предметної області

Основна задача сентимент аналізу тексту, що містить висловлення думок, може бути сформульована наступним чином: якщо надано текст, що є суб’єктивним висловлюванням, то за припущення, що висловлювання має єдиний об’єкт, виявити емоційний відтінок тексту як одну з двох полярностей: позитивну чи негативну (sentiment polarity).

Визначення, чи має наданий текст в цілому позитивне чи негативне забарвлення називається “класифікацією полярності настроїв” (sentiment polarity classification) або “класифікацією полярності” (polarity classification). Одним з мінусів даного підходу є те, що емоційну складову документа не завжди можна однозначно визначити, тобто документ може містити як ознаки позитивної оцінки, так і ознаки негативної [1].

В аналізі тональності тексту вважається, що текстова інформація ділиться на два типи: факти і думки. Ключовим поняттям є визначення думки.

Думки поділяються на два типи:

- проста думка;
- порівняння.

Проста думка містить висловлювання автора про один об'єкт. Вона може бути висловлена прямо: «Останні рішення ВТО просто прекрасні», або неявно: «Після важких і малоприємних реформ економіка почала зростати».

Думка першого типу може бути визначена формально: простою думкою називається кортеж з п'яти елементів (entity, feature, sentiment value, holder, time). В цьому визначенні автор (holder) висловив думку про аспект (feature) об’єкту entity в певний момент часу (time). Зазвичай виділяють два види емоцій (sentiment value): позитивна та негативна, тобто класифікація виконується за двома класами. Іноді додається третій - нейтральна думка. [2]

Часто об'єкт можне бути поданий у вигляді ієрархічної структури (рисунок 1.1). З кожною компонентою пов'язаний набір атрибутів. У наведеному вище визначенні думки, під аспектом мається на увазі і компоненти, і їх атрибути. Окремим випадком аспекту є сам об'єкт.

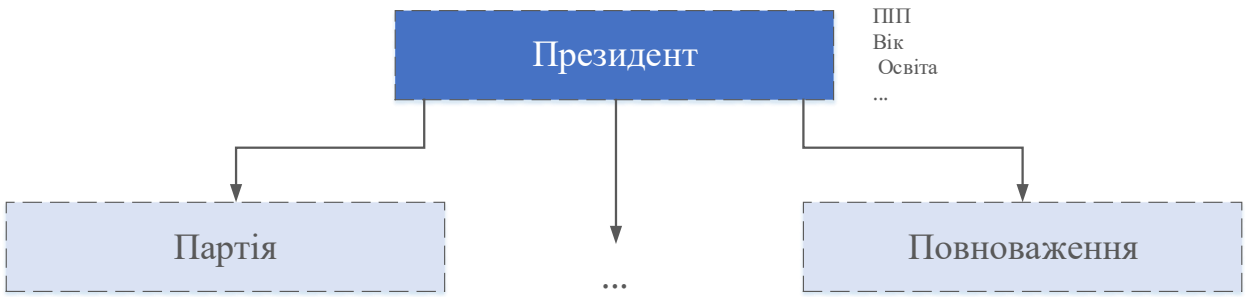

Рисунок 1.1 – Компоненти і атрибути об'єкту

Другий тип думок – порівняння – можна розділити на три види:

- порівняння аспектів об'єктів на користь одного (non–equal gradable);
- прирівнювання аспектів різних об'єктів (equative);
- перевага одного об'єкта над іншими (superlative).

Порівняння першого типу мають вигляд «аспект об'єкта 1 перевершує в чомусь аспект об'єкта 2», наприклад: «Ціни в інтернет-магазині «А» нижче, ніж в інтернет-магазині «В». Другий тип виражає схожість аспектів різних об'єктів, наприклад: «Ціни в інтернет-магазинах «А» та «Б» майже однакові». Прикладом третього типу може слугувати речення «В конкурсі на кращий магазин місяця магазин «А» перемагає магазин «Б».

Думка другого типу визначається як кортеж (Obj1, Obj2, A, holder, time). В даному кортежі Obj1 и Obj2 – множини порівнюваних за аспектом A об'єктів, які автор (holder) порівнює у момент часу time. На відміну від кортежу, який визначає думку першого типу, кортеж думки другого типу не містить прямої оцінки емоцій автора.

В аналізі тональності тексту часто зустрічається термін, пов'язаний з поняттям думки – суб'єктивність. Визначення об'єктивного і суб'єктивного речень наступне:

- об'єктивне речення відображає фактичну інформацію про що–небудь, тоді як суб'єктивне речення виражає чиїсь особисті почуття і припущення;
- об’єктивні речення зазвичай не мають емоційного забарвлення, тому аналіз тексту на наявність суб'єктивної інформації часто є підзадачею визначення полярності тексту. [3]

У комп'ютерній лінгвістиці текст природною мовою вважається неструктурованою інформацією. У завданнях, об'єднаних терміном «аналіз емоційного забарвлення тексту», визначається те, яким чином з тексту природною мовою витягується, аналізується і структурується інформація.

Отже, аналіз тональності тексту зазвичай включає в себе наступні основні завдання:

- визначення наявності емоційного забарвлення;
- визначення полярності тексту;
- вилучення аспектів з емоційно забарвленого тексту.

Задача визначення полярності тексту формулюється наступним чином: «визначити, яке емоційне забарвлення тексту, позитивне чи негативне?» Визначення полярності тексту зазвичай розглядається на декількох рівнях:

- на рівні документу. Основною задачею на цьому рівні є класифікація, чи повністю весь документ є відображенням позитивної чи негативної думки. Наприклад, для певної рецензії на товар, система автоматичного аналізу тональності тексту визначає, чи висловлює ця рецензія позивний настрій в цілому. Дана задача носить назву «класифікація полярності настроїв на рівні документу». Такий рівень деталізації передбачає, що кожен документ висловлює думку як єдину сутність. Тому він не може застосовуватись для документів, об’єктами яких є декілька сутностей. Однак в рамках данної задачі, зазвичай документи мають одне чітке емоційне забарвлення, бо коментарі до новин зазвичай є досить короткими емоційними текстами; [4]
- на рівні речення. На цьому рівні об’єктом дослідження є окреме речення. Проводиться аналіз чи висловлює певне речення в цілому позитивну чи

негативну думку. Аналіз на даному рівні близько пов’язаний з так званою «класифікацією суб’єктивності» (subjectivity classification), яка розрізняє речення (так звані об’єктивні речення) що висловлюють фактичну інформацію від речень, що висловлюють думки та погляди;

- на рівні сутності та аспекту. Обидва попередніх рівня не включають аналіз того, що саме сподобалося чи не сподобалося власнику думки. Аспектний рівень дозволяє виконати більш детальний аналіз. Замість того, щоб аналізувати лінгвістичні конструкції, аспектний рівень аналізує саму думку. Зазвичай, аналіз думки без аналізу її об’єкту має обмежене використання. Окрім того, визнання важливості аналізу об’єкту думки допомагає глибше зрозуміти проблему аналізу тональності. Наприклад, речення «Не зважаючи на поганий сервіс, мені все одно сподобався цей ресторан» має позитивний сентимент, але ми не можемо стверджувати, що воно є повністю позитивним. Власне, даний сентимент можна вважати позитивним лише в тому випадку, якщо в якості об’єкту обрано «ресторан». Якщо ж в якості об’єкту обрано «сервіс», то думка є повністю негативною. В багатьох дослідженнях об’єкти думки описуються сутностями та їх аспектами. Тобто, метою аналізу такого рівня є виявлення сентименту об’єкту та його властивостей. На такому рівні аналізу, можна отримати структурований підсумок думок щодо не тільки самого об’єкту, але і його властивостей, що перетворить неструктурований текст в структурований масив даних що може бути використаний для будь-яких типів аналізу.

Взагалі, аналіз на рівні документу та на рівні речення є досить складними, а аналіз на рівні аспектів є неймовірно важкою задачею.

Найбільш важливими індикаторами сентименту є «слова емоційного забарвлення» (sentiment words). Ці слова зазвичай використовуються для висловлення думки, позитивної чи негативної. Наприклад, «добре», «чудово», «неймовірно» – слова для висловлення позитивної думки, в той час як «погано», «жахливо», «сумно» – слова для висловлення негативної думки. Окрім безпосередньо слів, існують також

фрази та ідіоми. Слова та вирази є інструментарієм аналізу тональності з очевидних причин. Набір таких слів та виразів називається «лексичним словником».

Незважаючи на те, що слова та вирази для вираження емоційного забарвлення ж дуже важливими при аналізі тональності, просто використання їх не є досить ефективним. Проблема є комплексною та набагато складнішою. Нижче показані основні проблеми, з якими можна стикнутися при виконанні аналізу тональності за словником:

- позитивне чи негативне слово може приймати протилежний відтінок при використанні в іншій предметній області. Так, наприклад, «У цього фільму передбачуваний сюжет» є негативною характеристикою, а «У цього коду передбачувана поведінка» є позитивною;
- Наявність неологізмів та помилок у словах, а це досить поширене у інтернет-культурі явище, у багатьох випадках робить використання словникових методів недоцільним; [4]
- речення, що містить в собі слово емоційного забарвлення, може мати нейтральний сентимент. Цей феномен виникає зазвичай в декількох типах речень. Питальні речення та умовні речення є двома найважливішими типами, наприклад «Чи не могли в Ви порадити, яка з фотокамер Sony є найкращою?», та «Якщо я знайду дійсно хорошу камеру в цьому магазині, я її обов’язково куплю.» В обох цих реченнях є слова, що виражають позитивний настрій («найкраща», «хороша»), але жодне з цих речень не висловлює позитивну чи негативну думку щодо певної камери;
- речення, що містять сарказм з наявністю чи відсутністю слів емоційного забарвлення є дуже складними для аналізу, наприклад «Який чудовий телефон! Перестав працювати вже за два дні». Сарказм досить досить часто з’являється в політичних дебатах чи дискусіях, тому аналізувати політичні думки, як правило, досить важко. Ця проблема останнім часом має багато уваги, і навіть з’являються деякі практичні результати. [5]

Всі описані вище проблеми є досить серйозними труднощами для виконання аналізу тональності, що базується на лексичному словнику. Тому, методи, що використовують лексичний словник, як правило, застосовуються рідше, ніж альтернативні методи.

1.1.1 Класифікація думок під час аналізу тональності

В попередньому підрозділі ми визначили, що існують два типи думок – звичайні думки та порівняльні думки. Окрім такої класифікації, звичайні думки також розділяються на явні (explicit) та неявні (implicit) думки (або думки, що маються на увазі).

Звичайна думка. Звичайна думка дуже часто називається просто «думкою» в літературі, та має два основні під типи.

Явна думка. Явна думка – це думка, що була висловлена безпосередньо щодо самої сутності або аспекту цієї сутності. Наприклад: «Якість зйомки просто чудова».

Неявна думка. Неявна думка – це думка, що виражається неявно щодо сутності або аспекту, на основі ефекту, що має ця сутність на інші сутності. Такий підтип досить часто виникає в медичній області. Наприклад, речення «Після введення препарату я почуваюся гірше» висловлює неявну негативну думку щодо препарату, адже ця сутність «препарат» має негативній ефект на іншу сутність «самопочуття».

Більшість сучасних досліджень спираються на явні думки. Їх аналіз відбувається простіше. Аналізувати ж неявні думки досить складно. Наприклад, в області препаратів та медицині, необхідно знати, чи є наданий ефект бажаним чи небажаним. Наприклад, речення «Оскільки я дуже погано себе почуваю, лікар виписав мені цей препарат» не виражає негативної думки щодо препарату, адже «погано себе почуваю» сталося до моменту прийняття препарату.

Порівняльні думки. Порівняльна думка висловлює відношення подібності або відмінності між однією чи двома сутностями, та визначає, яку саме сутність врешті було обрано власником думки. Наприклад, речення «Кола на смак краще, ніж Пепсі» та «Кола найкраща» висловлюють дві порівняльні думки. Порівняльні думки, як правило, висловлюються в порівняльні формі прикметника або прислівника, хоча і не завжди (наприклад, «я надаю перевагу»).

Явна думка. Явна думка – це суб'єктивне твердження, що надає звичайну чи порівняльну думку, наприклад:

«Кола смакує добре» або «Кола смакує краще, ніж Пепсі»

Неявна думка. Неявна думка – це об'єктивне твердження, що має в собі звичайну чи порівняльну думку. Таке твердження, як правило, виражає бажаний чи небажаний факт, наприклад:

«Я купив цей матрац два тижні тому, і він деформувався», та «Заряд батареї телефонів Nokia кращий, ніж в телефонів Samsung».

Явні думки простіше виявити та класифікувати, ніж неявні. Більшість сучасних досліджень фокусується саме на явних думках. Досить мало досліджень було проведено щодо визначення та класифікації неявних думок.

1.1.2 Суб'єктивність та емоції

Існують дві концепції, що є дуже тісно пов'язаними з класифікацією емоційного забарвлення думок – суб'єктивність та емоції.

Об'єктивне речення визначає певну фактичну інформацію щодо навколишнього світу, в той час як суб'єктивне твердження висловлює почуття та думки окремої людини.

Прикладом об'єктивного твердження є наступне речення: «iPhone є продуктом компанії Apple». Прикладом суб'єктивного твердження є наступне речення: «Мені

подобається iPhone». Суб'єктивні твердження можуть виступати у самих різноманітних формах, таких як думки, твердження, бажання, переконання та інші. Існує певна проблема в літературі відрізнити суб'єктивність від висловлення думки. Під висловленням думки мається на увазі, що документ чи окреме речення висловлює позитивну чи негативну думку щодо певної сутності. Ці дві концепції не є еквівалентними, не дивлячись на те, що вони досить тісно перетинаються. Задача, що займається визначенням того, чи має документ суб'єктивне чи об'єктивне твердження носить назву «Класифікація суб'єктивності». Ми повинні мати на увазі наступне:

- суб'єктивне речення може не висловлювати ніякої думки. Наприклад, речення «Я думаю, що він пішов додому» є повністю суб'єктивним, але не висловлює нічого. Речення в попередньому прикладі також є повністю суб'єктивним та не висловлює позитивну, негативну чи нейтральну думку щодо камери або одного з її аспектів;
- об'єктивні речення можуть висловлювати думки на основі стверджування бажаних чи небажаних фактів. Наприклад, наступні два речення, що просто висловлюють факти об'єктивно також і висловлюють негативну думку (неявну негативну думку) щодо певних продуктів: «Ці навушники зламалися через два дні» або «Не дивлячись на те, що комп'ютер новий, він перестав вмикатися через місяць».

Окрім явних думок, які несуть в собі суб'єктивні висловлювання, дуже багато інших типів суб'єктивності вивчалися аналітиками, але не так інтенсивне. Дуже багато таких типів можуть також висловлювати ту чи іншу думку.

Емоції досить тісно пов'язані з висловленням думки. Вони були класифіковані та виділені в певні категорії – любов, здивування, радість, сум, страх. Сила емоційного забарвлення, що присутня в думці, як правило пов'язана з силою певної емоції, тому можна деколи зустріти не бінарну классифікацію на гарні/погані емоції а классифікацію за типами емоцій, що містить висловлювання. Найчастіше така классифікація відбувається за допомогою словникових методів, адже дуже важко

отримати достатньо велику вибірку для того щоб гарно розрізняти декілька видів емоцій у висловлюваннях.

1.1.3 Підзадачі аналізу тональності

Розглянемо модель сутності. Нехай сутність e_i представляється скінченним набором аспектів $A_i = \{a_{i1}, a_{i2}, ..., a_{in}\}$. E_i може бути виражена будь якім іншим набором виразів сутності $\{ee_{i1}, ee_{i2}, ..., ee_{in}\}$. Кожен аспект сутності може бути представлений як набір виразів $\{ae_{i1}, ae_{i2}, ..., ae_{in}\}$.

Також, нехай емоційно забарвлений документ d містить в собі думки щодо набору сутностей $\{e_1, e_2, ..., e_n\}$ та їх аспектів від певного набору власників думки в певний період часу.

Нарешті, маючи визначення емоційно забарвленого документу можна визначити основні задачі, що потрібно вирішити в процесі аналізу тональності.

Задача 1. Вилучення сутностей та їх категоризація. Вилучити всі вирази сутностей в емоційно забарвленому документі та виконати категоризацію та розбиття їх на класи сутностей. Кожен клас характеризує окрему сутність.

Задача 2. Вилучення аспектів сутностей та їх категоризація. Вилучити всі вирази сутностей в емоційно забарвленому документі та виконати категоризацію та розбиття їх на класи сутностей. Кожен клас характеризує окрему сутність.

Задача 3. Вилучення власників думки та їх класифікація. Вилучити власників думки та класифікувати їх. Ця задача є подібною до задач 1 та 2.

Задача 4. Вилучення часу та стандартизація. Необхідно вилучити всі часові проміжки, коли були висловлені думки їх власниками та стандартизувати різні формати представлення часу.

Задача 5. Класифікація емоційного забарвлення аспектів. Необхідно визначити, чи є думка що до аспекту a_{ij} позитивною, негативною чи нейтральною, або надати числове значення що визначає думку.

Кортеж з п'яти елементів, в вигляді якого ми дали визначення думці, надає дуже добре джерело для інформації, а також основу для генерації як якісних, так і кількісних підсумків. Загальна форма таких підсумків базується на аспектах і має назву підсумки, що базуються на аспектах.

Давайте наведемо приклад, як може виглядати підсумок аналізу тональності, основуючись на викладках, зроблених вище.

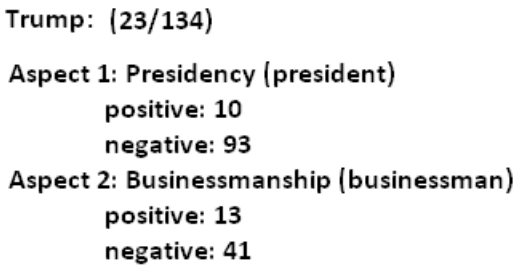

Рисунок 1.2 – Приклад підсумку аналізу тональності

В даному прикладі (рисунок 1.2) проводиться підсумок коментарів щодо Дональда Трампа. Цей підсумок є структурованим, на відміну від звичайного текстового підсумку. В даному прикладі, зверху ми бачимо висновок відносно всього об’єкта загалом. 23 відгуки мали позитивну думку щодо об’єкту і 134 - негативну. Президентство та бізнес – це два найпопулярніші в обговореннях аспекти сутності «Трамп». Наприклад, щодо першого 10 залишили позитивний коментар та 93 негативний. Кількість коментарів також може бути посиланням на власне самі коментарі. Маючи підсумок такого типу, можна швидко та зручно провести аналіз того, як користувачі висловлюються щодо всієї сутності в цілому, так її до окремих його компонентів. Якщо аналітик зацікавлений в певному компоненті сутності або додаткових деталях, він може скористатися посиланням та переглянути власне самі думки чи повні документи.

1.2 Існуючі розробки в області аналізу тональності

SentiStrength — система, розроблена M. Thelwall, K. Buckley, G. Paltoglou та D. Cai. Перш за все, дана система була розроблена для аналізу коротких структурованих неформальних текстів англійською мовою. Однак, вона може бути налаштована для роботи з текстами на ряді інших мов. Результат видається у вигляді двох оцінок – оцінка позитивної та негативної складової тексту (за шкалами від +1 до +5 та від -1 до -5 відповідно) Крім того, існує можливість надання оцінок у бінарній чи тернарній шкалах. Алгоритм шукає максимальне значення тональності в тексті для кожної шкали (як позитивної, так і негативної). При роботі алгоритм враховує принамні найпростішу взаємодію слів (наприклад, слова-підсилювачі посилюють значення тональності для слова, на яке вони діють) і ідіоматичні вирази. Недоліки системи: хоча система може бути налаштована для різних мов, але алгоритм не враховує специфіку кожної мови. Крім того, враховується лише загальна тональність тексту, без виділення аспектів. [4]

Ваал – система, розроблена В. Шалаком. У процесі своєї роботи система використовує перетворенні тексту в частотний словник, класифікацію деяких свів за психолінгвістичними категоріями. Недоліки системи: система не виконує аналіз семантики тексту, що веде до сильної обмеженості застосування продукту. Крім того, даний продукт є надто складним для простих користувачів, що не є фахівцями в області психолінгвістики.

RCO Fact Extractor – система, що була розроблена компанією RCO. Для роботи поданою системою використовується підхід, заснований на правилах. Враховується як синтаксична структура тексту, так і взаємодія різних типів слів. Робота компонента відбувається у п'ять етапів:

- розпізнавання всіх згадок про об'єкт у всіх формах, включаючи повні, короткі та інші форми згадок;

– фільтрування та синтаксичний розбір конструкцій, у яких знаходять відображення всі події і ознаки, що пов'язані з цільовим об'єктом;
– виділення і класифікація тих частин, де з найбільшою імовірністю виражається тональність, і тих пропозицій, які описують емоційно-коннотативні ситуації;
– для кожної пропозиції ухвалення рішення про класифікацію тональності на позитивну-негативну з урахуванням позиції, яку займають у її складі тональні і нейтральні слова та заперечення;
– підрахунок загальної тональності тексту на основі усіх оцінок тональностей пропозицій.

Для своєї роботи компонент використовує модулі синтаксичного аналізу тексту і ототожнення найменувань. Недоліки системи: відсутність кількісної оцінки тексту.

Brand Analytics – це система моніторингу та аналізу згадок в соціальних медіа (блогах, мікроблогах, соціальних мережах, форумах та ін.) і онлайн ЗМІ. Система має такі функції:

- відстежує обговорення в інтернеті компанії, її продукції, послуг, персоналу і конкурентів;
- автоматично аналізує знайдені згадки;
- надає звіти в режимі реального часу.

Brand Analytics призначена для маркетологів, PR–менеджерів, фахівців служби підтримки користувачів, менеджерів з продажу та всім, кому важлива думка користувачів про компанію та її послуги.

В системі Brand Analytics використовується модуль визначення тональності Eureka Engine, який може класифікувати російськомовні тексти за трьома видами – позитивні, негативні та нейтральні.

Модуль заснований на алгоритмі випадкових марковських полів з використанням тональних словників. Це дозволило досягти достатньої якості (середня точність складає близько 87%) і високої швидкості обробки текстів.

Навідміну від більшості інших реалізацій, поданий модуль дозволяє оцінити силу тональності. Таким чином, користувачу надається можливість не тільки отримати якісну емотивну оцінку документа в цілому щодо об'єкта, що цікавить тональності, а й кількісне співвідношення сили негативного та позитивного відношення до нього.

Модуль може працювати текстами новинного потоку, які написані відносно літературною мовою, так і «некласичною» мовою повідомлень у соціальних медіа.

YouScan – система для професійного моніторингу російськомовних соціальних медіа. YouScan виконує роль спостерігача для банків у соціальних мережах і дозволяє бути в курсі скарг і побажань споживачів, що публікуються в соціальних медіа; швидко реагувати і підвищувати якість обслуговування. Система також допомагає знайти нових клієнтів, які зацікавлені в банківських продуктах і послугах.

YouScan відстежує згадки банківських брендів, продуктів, конкурентів у блогах, форумах, соціальних мережах, і подає результати аналізу в зручному аналітичному вигляді, реалізую функції командної роботи. YouScan також дозволяє банкам знаходити в соціальних медіа гарячі ліди – потенційних клієнтів, користувачів, які зацікавлені в тих чи інших банківських продуктах або послугах, і передавати їх в CRM-систему банку.

1.3 Підходи до проведення аналізу тональності

1.3.1 Підходи до попередньої обробки даних

Як правило, перед тим, як починати саму класифікацію тональності, виконується попередня обробка даних. Етап попередньої обробки є таким ж важливим, як і етап безпосередньої класифікації, адже від того, в якому вигляді початковий документ буде представлений для методу аналізу, буде залежати точність. Для попередньої обробки коментарів автор пропонує використовувати

методи, що вже широко використовуються для обробки повідомленнь у соціальних мережах – видалення чисел, посиланнь на інтернет сторінки, власних імен, звертанння та слів з англомовного словника стоп-слів. В одночас, є сенс наголосити, що через те, що розглядаються саме англомовні коментарі, деякі процедури, наприклад стемінг, не мають такого вирішального значення, як наприклад для україномовних.

Другим використовуваним методом є Word2Vec, розробка компанії Google, що дозволяє описати слова як вектор чисел, певної розмірності. При цьому слова що часто зустрічаються на однакових позиціях, а отже імовірно мають схожий сенс будуть мати близьке векторне представлення, що отже зменшить складність задачі для SVM.

Загалом, мета цих процедур – якнайбільше зменшення розмірності задачі без втрати емоційної складової .

Окрім цих підходів, також часто виконується обробка заперечень.

1.3.1 Підходи до класифікації тональності

Для виконання безпосередньої класифікації також існує ряд методів. Всі ці методи відрізняються за точністю та швидкодією. До найбільш популярних методів відносять методи машинного навчання з учителем та без учителя, методи, основані на словниках та правилах, та ряд інших. Для побудови системи, що виконує автоматичний аналіз тональності, як правило використовуються методи машинного навчання. Методи машинного навчання без учителя, або навіть semi-supervised, як правильно, дають нижчу точність, ніж методи машинного навчання с учителем. Методи, основані на словниках та правилах дають непогану точність, але вони є дуже залежними від предметної області. Якщо існує необхідність виконати аналіз в декількох областях, необхідно складати декілька словників. Сам процес складання словника є досить важким, тому ці методи важко застосовувати для автоматичного

аналізу, що не залежить від предметної області. Лінгвістичний підхід може надати відносно точні результати, будучи реалізованим для наукових або журнальних статей або інших, граматично вірних текстів. Беручи до уваги той факт, що одне з головних застосувань аналізу тональності – бізнес-розвідка, стає зрозуміло, що інтернет-спільнота не може забезпечити дослідників граматично правильними текстами, або навіть текстами без орфографічних помилок. У зв'язку з цим, не варто і говорити про граматику і стилі письма рядових користувачів соціальних мереж. Окрім того, підхід, заснований на правилах, сильно прив'язаний до конкретної мови.

Висновки за розділом

В даному розділі проведено огляд загальної задачі аналізу тональності, надано визначення самому аналізу тональності та основним термінам, що використовуються при формулюванні даної проблеми.

Розглянуто, які додаткові задачі необхідно вирішити досліднику та проілюстровано, з якими основними проблемами стикаються дослідники при виконанні аналізу тональності. Також, описані можливі варіанти вирішення таких проблем.

Окрім того, описано, які наразі є існуючі дослідження в області аналізу тональності та їх програмні реалізації, а також проведена приблизна експертиза системи, що виконує автоматичний аналіз тональності на основі одного з можливих методів.

В якості підсумку, в даному розділі поставлені задачі дослідження даної магістерської дисертації.

2 АНАЛІЗ МЕТОДІВ ПОПЕРЕДНЬОЇ ОБРОБКИ ТА КЛАСИФІКАЦІЇ

Задача, що була описана в Розділі 1, постає як задача класифікації емоційно забарвлених думок на рівні документу, адже саме цілий документ є основою для аналізу. Досить велика кількість досліджень присвячена аналізу саме коментарів соціальних мереж. Тому, проблема, яка постає перед нами, може бути визначена наступним чином:

Визначення проблеми. Нехай надано документ d, що може бути представлений у вигляді $(\{e_i,\ s_i\}, h, t)$, де $\{e_i,\ s_i\}$ є парами обьект/сентимент відносно об'єкту, h – власник думки, t – час висловлювання думки. Визначити p(s) – полярність сентименту домументу, де сутність e, власник думки h та час висловлення думки t вважаються вже визначеними або не мають значення.

Існують дві проблеми, в залежності від того, яке значення приймає s. Якщо s приймає категоричне значення (позитивне чи негативне), то перед аналітиком постає проблема класифікації. Якщо s приймає чисельні значення в визначеному проміжку, постає проблема регресії.

Для того, щоб переконатися, що проблема має місце на практиці, існуючи дослідження роблять наступне неявне припущення.

Припущення: Класифікація або регресія сентименту припускає, що емоційно забарвлений документ d висловлює думку щодо єдиної сутності e та містить думки від єдиного власника думки d. Якщо документ достатньо короткий то зазвичай у ньому міститься лише одна пара $\{e_i,\ s_i\}$, або може бути поданий у такому вигляді (документ оцінює різні аспекти одного об'єкта, тож окремі сантименти відносно аспектів можно замінити на сантимент відносно об'єкта в цілому). [4]

На практиці, якщо емоційно забарвлений документ висловлює думки щодо більшої кількості сутностей, то ці думки можуть відрізнятися в залежності від сутностей. Наприклад, власник думки може мати позитивну думку щодо одних сутностей та негативну – щодо інших. Тому, зазвичай, не існує практичного сенсу

присвоювати всьому документу єдиний сентимент. Однак, для коротких текстів, таких як коментарі або твіти, у більшості випадків має місце відносно одного об’єкта, або декількох сутностей одного об’єкта. [4]. Проте, для форумів чи постів в блогах, таке припущення не є валідним, адже в такому випадку автор може висловлювати думки щодо багатьох сутностей.

2.1 Загальний підхід до вирішення задачі аналізу тональності

Більшість дослідників у своїх роботах покладаються на лінгвістичний підхід, і цьому є логічне пояснення. Алгоритми, засновані на правилах, дають більш точні результати, так як робота цих методів тісно пов'язана з семантикою слів, на відміну від методів машинного навчання, що оперують зі статистикою і теорією ймовірності. Але, як вже було зазначено, лінгвістичний підхід має ряд серйозних недоліків.

Лінгвістичний підхід може надати відносно точні результати, будучи реалізованим для наукових або журнальних статей або інших, граматично вірних текстів. Беручи до уваги той факт, що одне з головних застосувань аналізу тональності – бізнес-розвідка, стає зрозуміло, що інтернет-спільнота не може забезпечити дослідників граматично правильними текстами, або навіть текстами без орфографічних помилок. У зв'язку з цим, не варто і говорити про граматику і стилі письма рядових користувачів соціальних мереж. Окрім того, підхід, заснований на правилах, сильно прив'язаний до конкретної мови.

Методи, що основані на навчанні без вчителя, як правило мають невисоку точність, хоча і не потребує початковою навчаючої вибірки.

У зв'язку з вищесказаним, в цьому дослідженні розглядається підхід, заснований на методах машинного навчання з учителем, адже він є автоматизованим, дозволяє отримати необхідну точність та не є найскладнішим у реалізації.

Як правило, вирішення задачі аналізу тональності за допомогою методів машинного навчання з учителем розбивається на наступні етапи:

- етап 1. Вибір метрик для оцінки ефективності алгоритмів;
- етап 2. Вибір ознак, за якими буде відбуватися аналіз;
- етап 3. Вибір та реалізація алгоритмів класифікації;
- етап 4. Створення навчаючої вибірки;
- етап 5. Перевірка ефективності алгоритмів на основі обраних метрик.

2.2 Дослідження методів попередньої обробки даних

2.2.1 Представлення документу у вигляді вектору ознак

Якість результатів безпосередньо залежить від того, яким чином документ буде поданий для класифікатора, який набір характеристик буде використаний для складання вектору ознак.

Вектор ознак – це алгебраїчна модель подання текстів:

$$d_j = (w_{1j}, w_{2j}, \dots, w_{nj}), \tag{2.17}$$

де d_j – векторне представлення документу j, w_{1j} – вага терміну i в документі j, n – кількість термінів.

Найпоширенішим способом представлення документа в задачах аналізу природної мови– це у вигляді набору слів (bag-of-words) або набору N-грам. Наприклад, речення «Я люблю молочний шоколад» можна представити у вигляді набору уніграм (Я, люблю, молочний, шоколад) або біграм (Я люблю, люблю молочний, молочний шоколад).

Зазвичай, найкращі результати має використання уніграм та біграм, використання N-грам більш високих порядків (триграми і вище) призводить до втрати

ефективності, через те, що навчаюча вибірка в більшості випадків не має достатнього об’єму для використання N-грам вищих порядків. У більшості випадків має сенсоцінити результати із застосуванням уніграм, біграм і їх комбінації (Я, люблю, молочний, шоколад, Я люблю, люблю молочний, молочний шоколад). Залежно від типу даних використання уніграми може бути більш або менш доцільним за використання біграм. Також іноді комбінація уніграм і біграм дозволяє поліпшити результат.

Менш популярним способом є представлення тексту у вигляді символьних N-грам. Так, речення з прикладу може бути подане у вигляді наступних 4-символьних N-грам: «я лю», «люб», «юблю», «блю», і т.д. Поданий спосіб може здатися більш примітивним за попередні, адже набір символів фіксованої довжини не здається дуже інформативним, проте поданий метод за певних умовможе давати результати навіть краще ніж N-грами слів. При більш уважному вивченні можна побачити, що N-грами символів у якійсь мірі відповідні морфемам слів, наприклад у випадко слова «люблю» корінь «люб» несе в собі його зміст. Символьні N-грами можуть бути корисні в двох випадках:

- за умови великої кількості орфографічних помилок у тексті – набір символів у тексті з помилками буде майже однаковим з набором символів у тексті без помилок;
- для мов зі змінною морфологією (наприклад, для української) – слова можуть мати багато різних варіацій, але при цьому корінь слова, а отже і набір символів, залишається незмінним.

Хоча символьні N-грами застосовуються менш часто за N-грами слів, але часом їх використання може покращити результати.

Чаом можуть бути використані також і інші ознаки: частини мови, наявність в тексті смайлів, знаків оклику, запереченя, вигуки і т.д.

Для повноцінного розуміння принципу векторної моделі тексту, необхідно пояснити, як саме розраховуються ваги векторів. Існує кілька базових функцій ваг. В інформаційному пошуку найбільш поширеним методом оцінки ваги ознак є TF-IDF.

TF-IDF (term frequency, inverse document frequency) – статистична міра, яка використовується для оцінки важливості слова в контексті документу, що є частиною колекції документів.

TF (term frequency – частота слова) – відношення числа входження деякого слова до загальної кількості слів документа. Поданий компонент показує частоту, а отже можна припустити, що і важливість слова в межах окремого документа.

IDF (inverse document frequency – зворотна частота документа) – інверсія частоти, з якою деяке слово зустрічається у інших документах. Облік IDF зменшує вагу часто і широко вживаних слів.

Отже, міра TF-IDF утворюється з двох співмножників: TF і IDF.

Для аналізу тональності цей метод не дає хороших результатів. Причиною цього є те, що для аналізу тональності не так важливі слова, які часто повторюються в тексті (тобто слова з високим TF), на відміну від завдання пошуку.

Було виявлено, що бінарна функція зважування векторів більш ефективна. [6] Це означає, що наявність терміну в документі важливіше, ніж його частота. Бінарні вектори представлені як послідовність нулів і одиниць: якщо конкретний термін зі словника вибірки зустрічається в тексті – вага терміну буде дорівнює 1, інакше – 0. Частотні вектори формуються на основі кількості входжень певного терміну в класі документів.

Наприклад, речення «Я люблю молочний шоколад» буде представлене у вигляді наступного вектору (ми опускаємо слова з вагою = 0):

{ " Я ": 1 , " люблю ": 1 , " молочний ": 1 , " шоколад ": 1 }

Однак, існують методи оцінки важливості слів, які обчислюють ваги слів, що дають набагато кращі результати при класифікації тональності, наприклад, дельта TF-IDF .

Ідея методу дельта TF-IDF полягає в тому, щоб дати більшу вагу для слів, які мають не-нейтральну тональність, тому що саме такі слова визначають тональність всього тексту. Формула для розрахунку ваги слова *w* наступна:

$$V_{t,d} = C_{t,d} * \log\left(\frac{|N| * P_t}{|P| * N_t}\right), \quad (2.18)$$

де:

- $V_{t,d}$ – вага слова t в документі d;
- $C_{t,d}$ – кількість разів, скільки слово t зустрічається в документі d;
- $|P|$ – кількість документів з позитивною тональністю;
- $|N|$ – кількість документів з негативною тональністю;
- P_t – кількість позитивних документів, де зустрічається слово t;
- N_t – кількість негативних документів, де зустрічається слово t.

Припустимо, ми працюємо з колекцією відгуків фільмів. Розглянемо три слова: «відмінний», «нудний», «сценарій». Найголовніше у формулі дельта TF-IDF – це другий множник log (...). Саме він буде різний у цих трьох слів:

- слово «відмінний» швидше за все зустрічається в більшості позитивних (P_t) відгуків і майже не зустрічається в негативних (N_t), в результаті вага буде великим позитивним числом, тому що відношення P_t/N_tбуде числом набагато більше 1;
- слово «нудний» навпаки зустрічається в основному в негативних відгуках, тому ставлення P_t/N_t буде менше одиниці і в підсумку логарифм буде негативним. В результаті вага слова буде негативним числом, але великим по модулю;
- слово «сценарій» може зустрічатися з однаковою ймовірністю і в позитивних, так і в негативних відгуках, тому ставлення P_t/N_t буде дуже близько до одиниці, і в підсумку логарифм буде близький до нуля. Вага слова буде практично дорівнює нулю.

В результаті вага слів з позитивною тональністю буде великим позитивним числом, вага слів з негативною тональністю буде негативним числом, вага нейтральних слів буде близькою до нуля. Таке зважування вектору ознак в більшості випадків дозволяє поліпшити точність класифікації тональності.

2.2.2 Стемінг та лематизація

У деяких дослідженнях (перш за все у дослідженнях не-англомовних текстів) при поданні тексту все слова проходять через процедуру стемінгу (видалення закінчення) або лематизації (приведення до початкової формі). Мета цієї процедури – зменшення розмірності задачі, іншими словами – якщо в тексті зустрічаються однакові слова, але з різними закінченнями, за допомогою стемінгу і лематизації можна їх привести до одного виду.

Однак, на практиці це зазвичай не дає ніяких відчутних результатів. Причина цього в тому, що, позбавляючись від закінчень слів, ми втрачаємо морфологічну інформацію, яка може бути корисна для аналізу тональності. Наприклад, слова «хочу» і «хотів» мають різну тональність. Якщо в першому випадку тональність швидше за все позитивна, тому що автор може висловлювати надію і позитивні емоції, то у дієслова в минулому часі, тональність може бути негативною, якщо автор висловлює жаль.

2.3 Дослідження методів класифікації тональності

2.3.1 Методи навчання з учителем

Задача аналізу тональності, як правило, формулюється як задача класифікації на два класи – позитивний та негативний. Дані, що використовуються для навчання та тестування, це, як правило, відгуки на певні товари. Оскільки онлайн відгуки, в більшості, надають також і чисельну оцінку продукту (як правило, в проміжку 1-5), класи «позитивний» та «негативний» визначаються, використовуючи цю ж саму рейтингову систему. Наприклад, відгук, що містить 4 або 5 зірочок відноситься до класу «позитивний», а всі інші, з оцінками 1 або 2 – до класу «негативний». Більшість

досліджень не використовують клас «нейтральний», що значно спрощує задачу класифікації. Однак, з рейтинговою системою можливе використання цього класу. До нього відносяться відгуки з оцінкою 3.

Задача аналізу тональності є, по свої суті, задачею текстової класифікації. Традиційна текстова класифікація, як правило, класифікує текст за різними темами, наприклад політика, наука, спорт на інші. В такому типі класифікації, ключовими компонентами є певні слова, що відносяться до певної теми. Однак, в задачі аналізу тональності, емоційно забарвлені слова, або слова сентименту, що визначають позитивну чи негативну думку є більш важливими, наприклад чудово, прекрасно, погано, жахливо.

Оскільки аналіз тональності є задачею текстової класифікації, для її вирішення може бути застосований будь-який метод, що застосовується для текстової класифікації, наприклад, класифікація методом наївного Баєса, метод максимум ентропії або методом опорних векторів (SVM, support vector machines). Вперше такі методі були використані для аналізу тональності в роботі Pang, Lee and Vaithyanathan (2002). В даній роботі було показано, що використовуючи уніграми (набори слів) як аспекти класифікації як з методом наївного Баєса, так і з методом опорних векторів. [36]

В наступних дослідженнях, дуже багато інших алгоритмів було застосовано великою кількістю аналітиків. Так само, як і інші додатки, що використовують машинне навчання з учителем для класифікації, ключовою задачею аналізу тональності є визначити набір ознак (features). Прикладом таких ознак можуть бути наступні:

- терміни та частота їх використання. Ці ознаки є окремими словами (уніграмами) та їх n-грам, та асоційованих з ними кількісних параметрів частоти використання. Ці ознаки є найбільш часто використовуваними в традиційній текстовій класифікації. В деяких випадках, може також враховуватись і порядок слів у реченні. Схема TF-IDF, що присвоює кожній ознаці певну вагу, також використовується досить часто. Як і в

традиційній текстовій класифікації, ці ознаки показали себе неймовірно ефективними і при аналізі тональності;

- частини мови. Частина мови (POS, Part Of Speech) кожного слова також може бути досить важливою. Слова з різними частинами мови можуть бути проаналізовані по-різному. Наприклад, було показано, що прикметники є дуже важливими при визначенні думок та в процесу аналізу тональності. Тому, деякі аналітики розробили окремий підхід до аналізу прикметників як спеціальних ознак. Однак, можливим також є використання всіх POS-тегів для всіх n-грам як ознак. POS-теги, або теги частини мови, показані в табл. 2.1;
- емоційно забарвлені слова та вирази. Емоційно забарвлені слова – це такі слова в тексті, що використовуються для висловлювання позитивної чи негативної думки. Наприклад, слово хороший використовуються для висловлення позитивною думки, в той час як слово поганий, як правило, використовується для висловлення негативної думки. Більшість емоційно забарвлених слів є прикметниками чи прислівниками, але інколи зустрічаються і іменники (дурниця, нісенітниця тощо). Окрім окремих слів, іноді використовуються повні фрази або вирази, ідіоми, що визначають певний сентимент, наприклад англійський вираз «cost someone an arm and leg», тобто коштувало дуже багато грошей;
- правила думок. Окрім емоційно забарвлених слів та виразів, існують також дуже багато інших виразів та мовних сполучень, що можуть бути використаними для вираження споживацьких думок;
- зворотні настрої. Такі вирази використовуються для того, щоб змінити полярність висловленої думки. Тобто, змінити полярність думки з позитивної на негативну чи навпаки. Найбільш часто в якості такої ознаки використовуються заперечення. Наприклад, вираз «Мені не подобається ця камера» є негативним. Існують також інші типи таких ознак, але їх потрібно аналізувати обережно, адже не завжди поява такої ознаки

призводить до того, що полярність змінюється. Наприклад, слово «не» в виразі «не тільки, але також» не змінює полярність;

– синтаксична залежність. Ознаки, що основані на залежностях, також були випробувані аналітиками при аналізі тональності за допомогою методів машинного навчання.

Навчання з учителем – способ машинного навчання, в ході якого система навчається за прикладами «стимул-реакція». З погляду кібернетики, є одним з видів кібернетичного експерименту. Між входами і еталонними виходами (стимул-реакція) може існувати деяка залежність, але вона невідома. За відомою сукупністю прецедентів – пар «стимул-реакція», що називається навчальною вибіркою, потрібно відновити залежність (побудувати модель, входом якої буде стимол, а виходом – реакція, що буде придатна для прогнозування). Для вимірювання точності відповідей, так само як і в навчанні на прикладах, може вводитися функціонал якості.

Схему експерименту можна побачити нижче (рисунок 2.1-2.2).

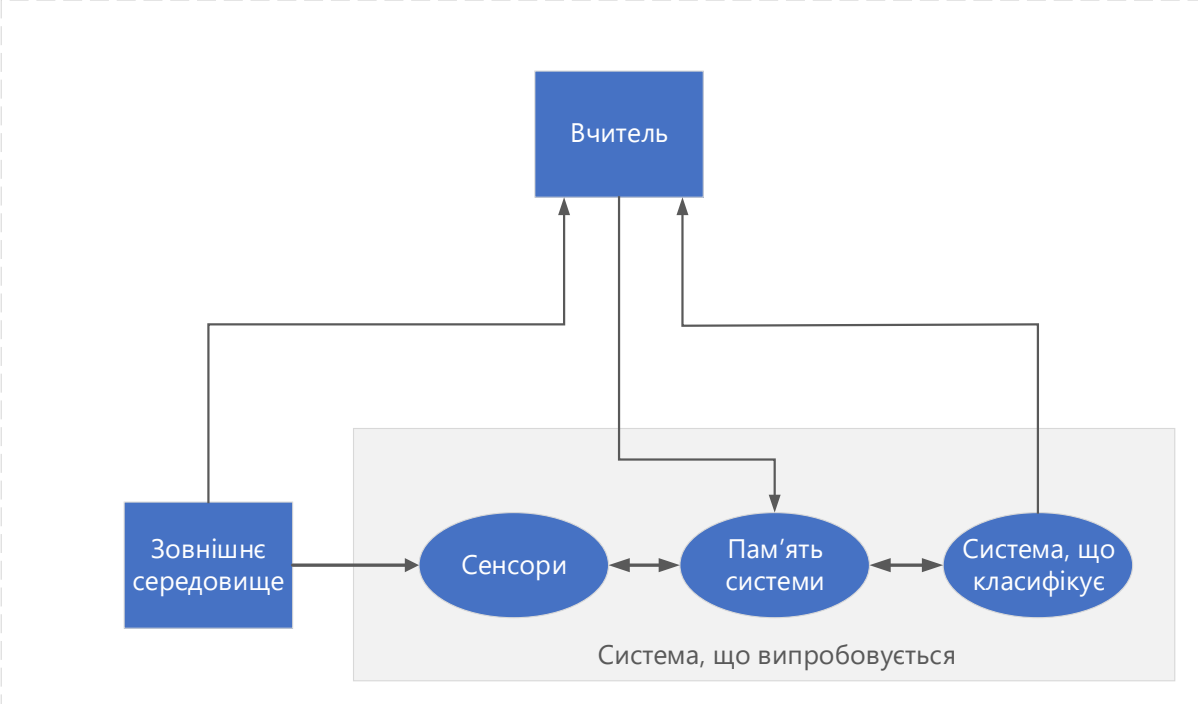

Рисунок 2.1 – Проста експериментальна схема

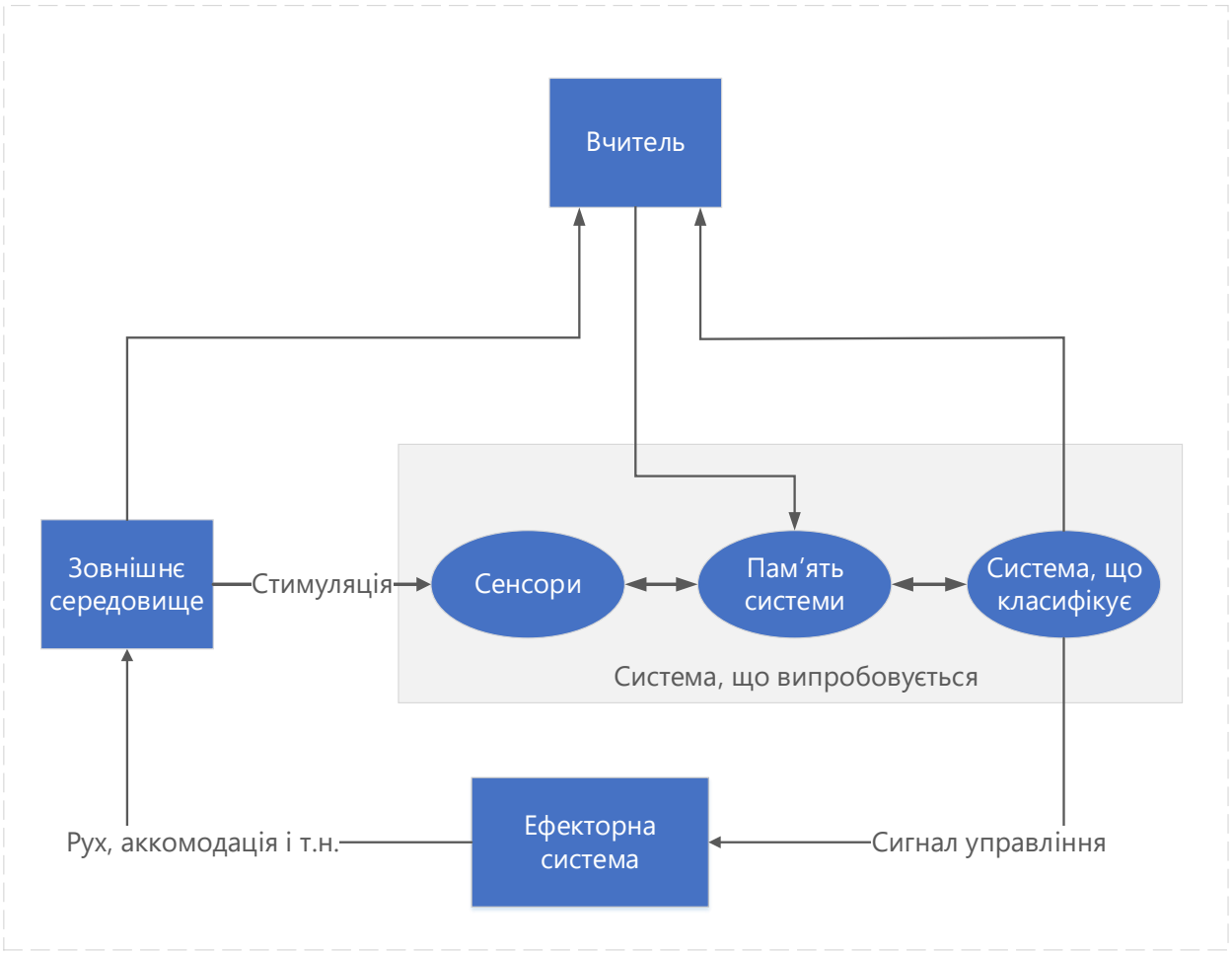

Рисунок 2.2 – Загальна експериментальна схема

Даний експеримент (рисунок 2.2) являє собою окремий випадок експерименту зі зворотним зв'язком. Постановка даного експерименту припускає наявність експериментальної системи, методу навчання і методу випробування системи або вимірювання характеристик.

Для аналізу тональності тексту необхідно вирішити задачу класифікації, що являє собою системний розподіл досліджуваних предметів чи явищ за видами або типами, за якими-небудь істотними ознаками, та, можливо, розташування їх у певному порядку, що відбиває ступінь цієї схожості.

Задача класифікації – формалізована задача, яка полягає у віднесення містить множиний об'єктів до певних класів. Нехай задана кінцева множина об'єктів та класів до яких відноситься кожен з поданих об’єктів (вибірка). Нехай також невідомо до

якого класу належать інші надані об’єкти. Метою задачі є створення алгоритму, що буде здатний до класифікації довільного об'єкту з вихідної множини.

У машинному навчанні завдання класифікації вирішується, як правило, за допомогою методів штучної нейронної мережі при постановці експерименту у вигляді навчання з учителем.

Наївний класифікатор Баєса – простий імовірнісний класифікатор, заснований на застосуванні Теореми Баєса зі строгими (наївними) припущеннями про незалежність.

Перевагою цього підходу є те, що вимоги до розміру вибірки скорочуються від експоненційних до лінійних. Недоліком – те, що модель може бути точною лише у випадку, коли виконується припущення про незалежність, інакше, строго кажучи, обчислені ймовірності вже не є точними. Однак часто результати роботи класифікатора продовжують корелювати з істинною приналежністю образів до класів навіть за умови істотної залежності.

Імовірнісна модель для класифікатора – це умовна модель

$$p(C|F_1, \dots, F_n) \tag{2.1}$$

від залежної змінної класу C з малою кількістю результатів чи класів, залежна від кількох змінних $F_1 \quad F_n$. Проблема полягає в тому, що коли кількість властивостей n у рівнянні (2.1) дуже велика чи коли властивість може приймати велику кількість значень, коли будувати таку модель на імовірнісних таблицях стає неможливо. Тому ми переформулюємо модель, щоб зробити її легко оброблюваною.

Використовуючи теорему Баєса, запишемо

$$p(C|F_1, \dots, F_n) = \frac{p(C)p(F_1, \dots, F_n|C)}{p(F_1, \dots, F_n)} \tag{2.2}$$

На практиці інтерес викликає лише чисельник цього дробу, оскільки знаменник не залежить від C і значення властивостей F_i дані, тому знаменник – константа. Чисельник еквівалентний сумісній імовірності моделі

$$p(C, F_1, \dots, F_n) \tag{2.3}$$

яка може бути переписана наступним чином, використовуючи повторно додатки визначень умовної імовірності:

$$\begin{aligned} &p(C, F_1, \dots, F_n) = \\ &p(C)p(F_1, \dots, F_n|C) = \\ &p(C)p(F_1|C)p(F_2, \dots, F_n|C, F_1) = \\ &p(C)p(F_1|C)p(F_2|C) \dots p(F_n|C, F_1, \dots, F_{n-1}) \end{aligned} \tag{2.4}$$

і так далі. Тепер можна використовувати «наївні» припущення умовної незалежності: припустимо, що кожна властивість F_i умовно незалежна від властивості F_j при $j \neq i$. Це означає:

$$p\left(F_i \middle| C, F_j\right) = p(F_i|C), \tag{2.5}$$

таким чином, сумісна модель може бути виражена як:

$$p(C|F_1, \dots, F_n) = p(C)p(F_1|C)p(F_2|C) \dots = p(C)\prod_{i=1}^{n} p(F_i|C) \tag{2.6}$$

Це означає, що з припущення про незалежність, умовний розподіл по класовій змінній C може бути виражено так: де Z – це масштабний множник,

$$p(C|F_1, \dots, F_n) = \frac{1}{Z} p(C) \prod_{i=1}^{n} p(F_i|C) \quad (2.7)$$

де Z – це масштабний множник, що залежить тільки від $F_1, \dots, F_n$, тобто константа, якщо значення змінних відомі.

До переваг цього класифікатора можна віднести наступні:

- простота реалізації;
- досить швидкий процес навчання;
- незважаючи на те, що припущення про незалежність класифікаційних ознак не є вірним в природній мові (значення слова залежать від контексту), НБК часто показує хороші результати при текстовій класифікації.

До недоліків цього методу можна віднести наступні:

– значення, що повертаються при класифікації, не можна трактувати, як вірогіднісні. Таким чином не можна відповісти на питання, з якою часткою впевненості вийшов результуючий клас;
– так як в природній мові слова не є незалежними, НБК не є найоптимальнішім методом.

Метод опорних векторів відноситься до сімейства лінійних класифікаторів. Метою лінійної класифікації є пошук гіперплощини просторі ознак, що розділяє всі об'єкти на два класи. Основна ідея методу опорних векторів полягає в пошуку роздільної гіперплощини, максимально віддаленої від найближчих до неї точок в просторі ознак.

Даний метод в своїй більшості використовується для задач бінарної класифікації. Однак, існує також можливість застосування його для класифікації на більшу кількість класів.

Основну ідею методу можна проілюструвати на наступному прикладі: нехай на площині дані точки, що належать до навчаючої вибірки та можуть бути розділені на

два класи. Можна провести пряму лінію, що буде розділяти ці два класи (рис.). Далі, всі нові точки, що належать вже до тестової вибірки, будуть класифіковані наступним чином: точка, що знаходиться вище прямої потрапляє до класу A, а точка, що знаходиться нижче прямої, потрапляє до класу B.

Така пряма маж назву розділяючої прямої. Однак, в просторах вищих розмірностей пряма вже не буде розділяти ці класи, так як поняття «нижче прямої» або «вище прямої» втрачає будь-який сенс. Тому замість прямих необхідно розглядати гіперплощини – простори, розмірність яких на одиницю менше, ніж розмірність початкового простору. Для простору $\mathbb{R}^3$, наприклад, гіперплощина – це звичайна двомірна площина.

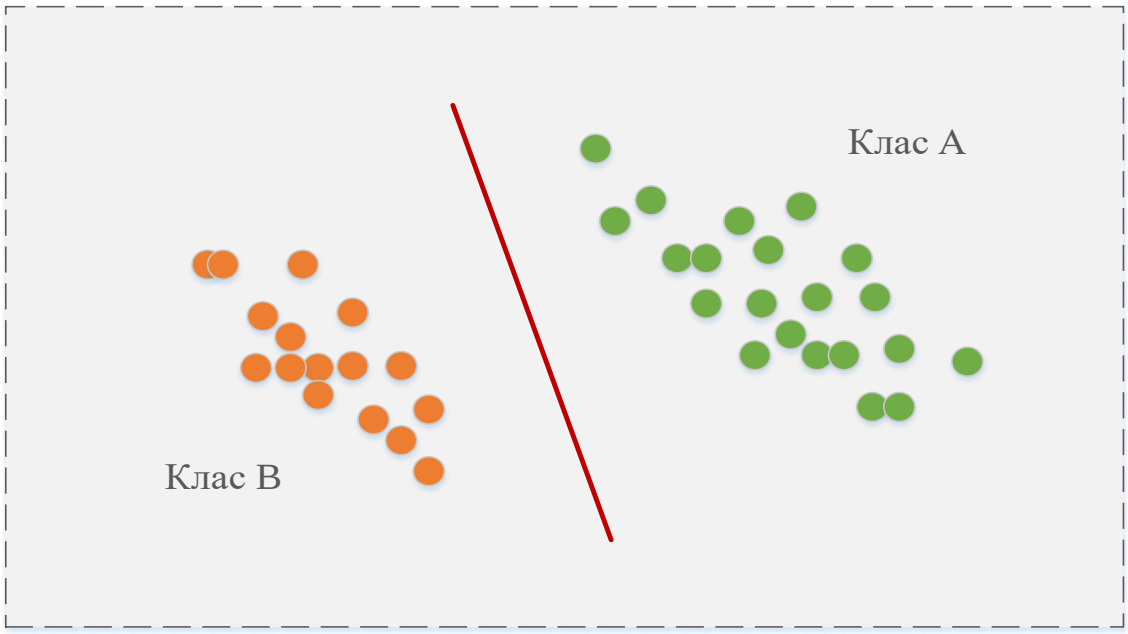

Рисунок 2.3 – Класифікація на два класи за допомогою роздільної прямої

Для прикладу, наведеного на рисинку 2.3 існують декілька прямих, що розділяють два класи, як це показано на рисунку 2.4.

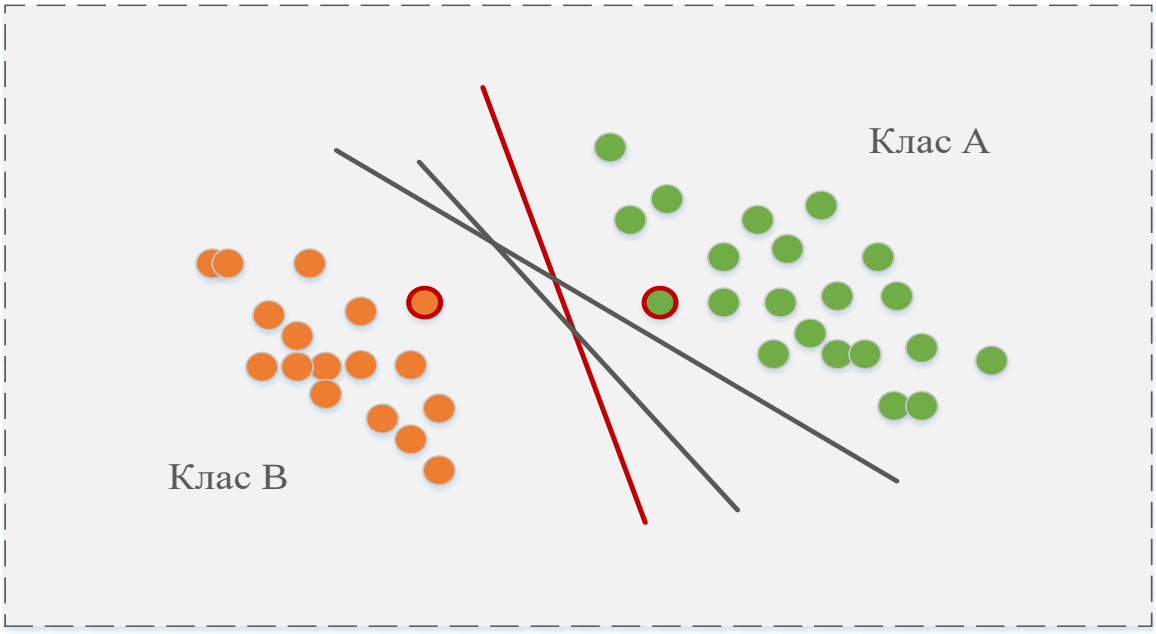

Рисунок 2.4 – Різні роздільні прямі

З точки зору точності класифікації найкраще вибрати пряму, відстань від якої до кожного класу є максимальною. Іншими словами, виберемо ту пряму, яка розділяє класи найкращим чином (рисунок 2.3). Така пряма, а в загальному випадку – гіперплощина, називається оптимальною розділяючої гіперплощиною.

Вектори, що лежать ближче всіх до розділяючої гіперплощини, називаються опорними векторами (support vectors). На рис. 2 вони позначені червоним кольором.

Розглянемо математичне підґрунтя цього методу.

Нехай маємо таку навчаючу вибірку:

$$(x_1 y_1), \dots, (x_m y_m), \qquad x_i \epsilon \mathbb{R}^n, \qquad y_i \epsilon (-1; 1) \tag{2.8}$$

Метод опорних векторів побудує класифікуючу функцію F у вигляді:

$$F(x) = sign(< w, x > + b) \tag{2.9}$$

де:

- $<,>$ – скалярний добуток;
- w – нормальний вектор, що розділяє гіперплощини;
- b – допоміжний параметр.

То об’єкти, для яких $F(x) = 1$ потрапляють в один клас, а ті об’єкти, для яких $F(x) = -1$, потрапляють в інший клас. Вибір саме такої функції невипадковий: будь-яка гіперплощина може бути задана у вигляді $< w, x > +b = 0$ для деяких w та b.

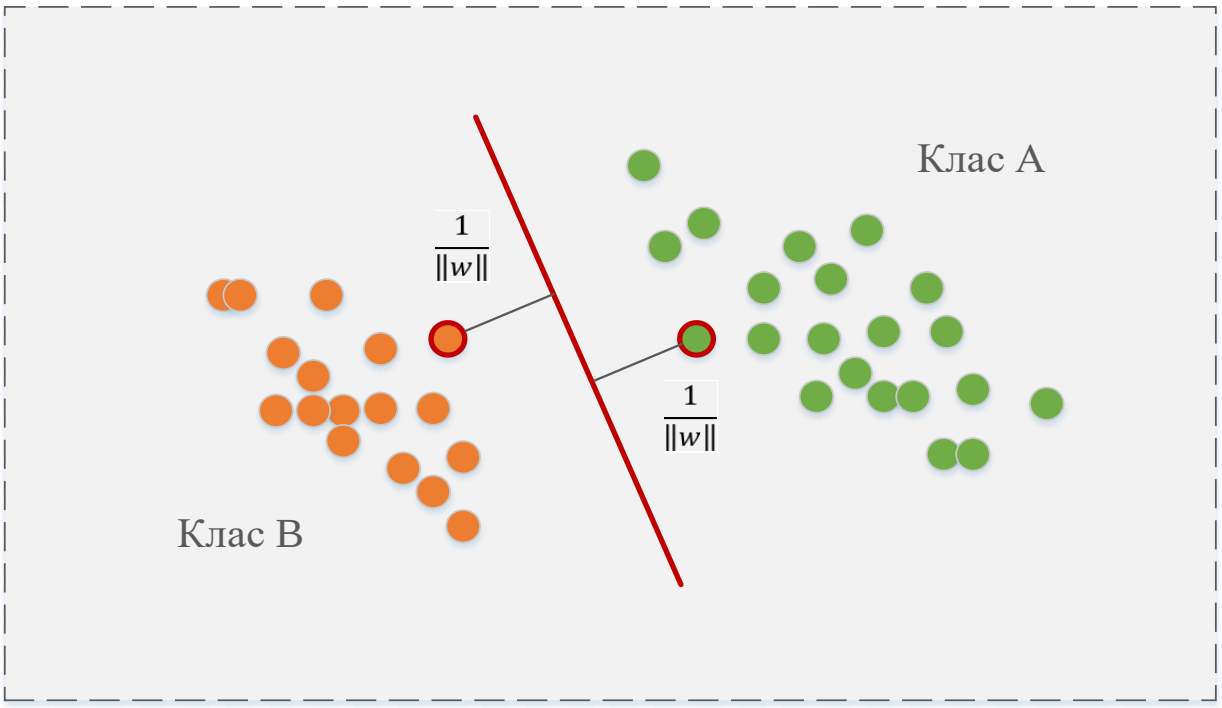

Рисунок 2.5 – Знаходження максимуму відстані

Далі, ми хочемо вибрати такі w та b, які максимізують відстань до кожного класу. Можна підрахувати, що дана відстань дорівнюватиме $\frac{1}{\|w\|}$ (рисунок 2.5). Проблема знаходження максимуму $\frac{1}{\|w\|}$ еквівалентна проблемі знаходження максимуму $\|w\|^2$. Запишемо все це у вигляді системи та задачі оптимізації.

$$\begin{cases} \arg\min\limits_{w,b} \|w\|^2 \\ y_i(< w, x_i > +b)) \geq 1, i = 1, \dots, m \end{cases} \tag{2.9}$$

Ця задача є стандартною задачею квадратичного програмування і вирішується за допомогою множників Лагранжа.

На практиці, випадки, коли дані можна розділити гіперплощиною, або, як ще кажуть, лінійно, досить рідкісні. Приклад лінійної нероздільності можна побачити нижче (рисунок 2.6).

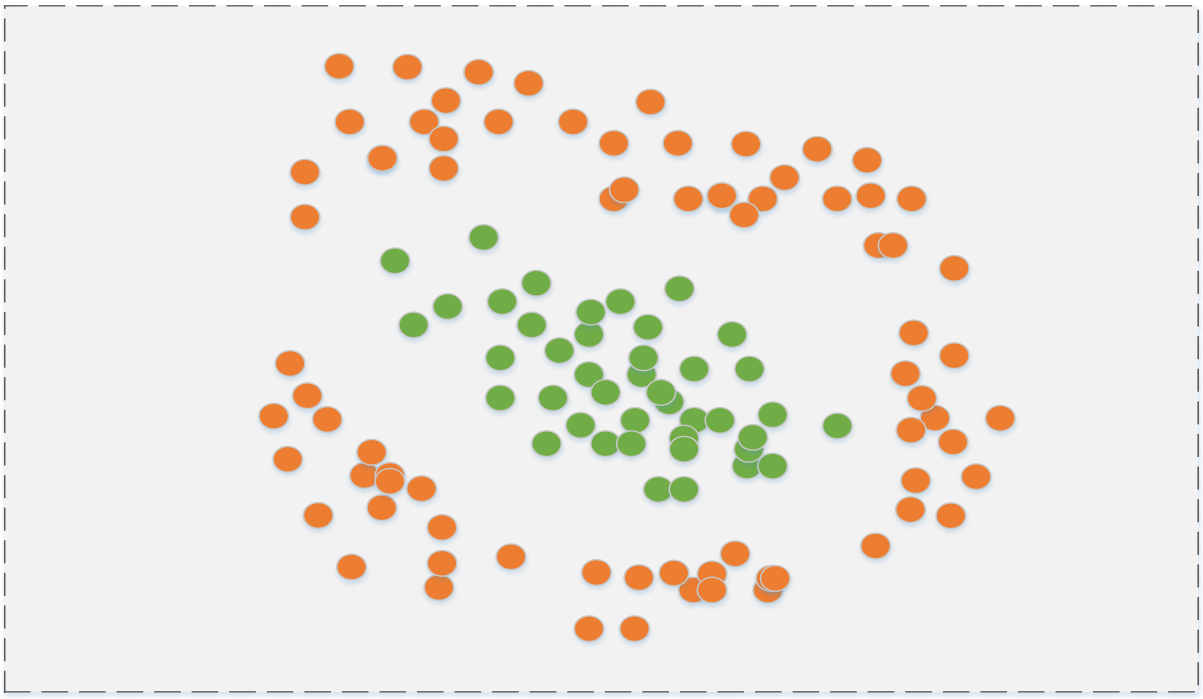

Рисунок 2.6 –Лінійно нероздільні класи

В цьому випадку поступають так: всі елементи навчальної вибірки вкладаються в простір X більш високої розмірності з допомогою спеціального відображення $\varphi: \mathbb{R}^n \rightarrow X$. При цьому відображення φ вибирається так, щоб в новому просторі X вибірна була лінійно роздільна.

Класифікуюча функція F приймає наступний вигляд:

$$F(x) = sign(< w, \varphi(x) > + b) \tag{2.10}$$

Вираз $k(x, x') = < \varphi(x), \varphi(x') >$ називається ядром класифікатора. математичної точки зору ядром може служити будь-яка позитивно визначена симетрична функція двох змінних. Позитивна визначеність необхідна для того, щоб

відповідна функція Лагранжа в задачі оптимізації була обмежена знизу, тобто задача оптимізації була б коректно визначена. Точність класифікатора залежить, зокрема, від вибору ядра. Найчастіше на практиці зустрічаються такі ядра, як, наприклад, поліноміальне:

$$k(x, x') = (< x, x' > + const)^d \quad (2.11)$$

2.3.2 Методи навчання без учителя

Оскільки емоційно забарвлені слова (або слова-сентименти) є домінуючим фактором для аналізу тональності, не важко уявити, що емоційно забарвлені слова та вирази можуть бути використані в методах для аналізу тональності з відсутністю вчителя. Подібні методи були описані в роботі Turney [7]. Такий метод виконує класифікацію, що основана на певних синтаксичних структурах, які як правило використовуються людьми для висловлення думок. Такі синтаксичні структури, або шаблони, як правило, базуються на певних частинах мови (Part-of-speech, POS) та їх тегах. Алгоритм для такої класифікації складається з наступних кроків:

- крок 1. Два послідовних слова вилучаються, якщо їх POS теги відповідають одному з шаблонів, що наведені в табл. 2.1. Наприклад, шаблон 2 означає, що 2 послідовних слова вилучаються, якщо перше слово – прислівник, а друге слово – прикметник. Третє слово, що не вилучається, це – іменник. Наприклад, для такого речення «Це піаніно має такий чудовий звук», «чудовий звук» вилучається, адже ця фраза відповідає шаблону 1, де перше слово – це прикметник, а друге слово – іменник. Причина, по якій використовуються ці шаблони, це те, що слова JJ, RB, RBR та RBS як правило висловлюють певну думку. Іменники та дієслова виступають у ролі контексту, тому що саме в залежності від контексту JJ, RB, RBR та RBS

можуть висловлювати різні думки, позитивні чи негативні. Так, наприклад, прикметник (JJ) «непередбачуваний» буде висловлювати негативну думку, якщо це відгук про машини («непередбачуване рульове управління»), або позитивну думку, якщо мова йде про відгуки до фільмів («непередбачуваний сюжет»).

- крок 2. Необхідно оцінити орієнтацію емоційного забарвлення (або сентименту) всіх вилучених фраз, використовуючи метрику точної взаємної інформації (Pointwise mutual information, PMI):

$$PMI(term_1, term_2) = \log_2\left(\frac{\Pr(term_1 \wedge term_2)}{\Pr(term_1)\Pr(term_2)}\right) \quad (2.12)$$

PMI вимірює ступінь статистичної залежності між двома термінами. Тут, $\Pr(term_1 \wedge term_2)$ – це фактична ймовірність спільної появи терму 1 та терму 2, а $\Pr(term_1)\Pr(term_2)$ – спільна ймовірність виникнення двох доданків, якщо вони є статистично незалежними. Емоційне забарвлення фрази (позитивне чи негативне) вираховується з огляду на його зв'язок з позитивним опорним словом "відмінно" і негативним опорним словом "погано":

$$SO(\text{фраза}) = PMI(\text{фраза, відмінно}) - PMI(\text{фраза, погано}) \quad (2.13)$$

Ймовірності обчислюються на основі запитів до пошукової системи, і збираючи кількість влучень (hits). Для кожного запиту, пошуковий механізм видає набір релевантних документів, що і є кількістю влучень. Таким чином, виконуючи пошук по двом термам разом та окремо, ймовірності в формулі можуть бути прораховані. В роботі Tuney [7], був використаний пошуковий двигун AltaVista, тому що він має оператор ПОРЯД, щоб обмежити пошук документами, які містять слова в межах десяти слів один від одного в

довільному порядку. Нехай hits – це отримана кількість влучень. Тоді рівняння (2.12) можна переписати наступним чином:

$$SO(\text{фраза}) = \log_2\left(\frac{hits(\text{фраза ПОРЯД відмінно})\, hits(\text{погано})}{hits(\text{фраза ПОРЯД погано})\, hits(\text{відмінно})}\right) \quad (2.14)$$

- крок 3. Маючи певний огляд, алгоритм обчислює середнє значення SO всіх фраз в огляді і класифікує огляд, як позитивний результат, якщо середній SO є позитивним і як негативний – в іншому випадку.

Підсумкова класифікація по точності оглядів з різних областей становить діапазон від 84% для автомобільних оглядів до 66% для оглядів фільмів.

2.3.3 Методи, засновані на словниках

Іншим методом є метод, оснований на словниках. Цей метод використовує словник емоційно забарвлених слів та фраз. Кожному слову чи фразі привласнена полярність та сила, і включає в себе інтенсифікацію і заперечення, щоб обчислити емоційне забарвлення кожного документу. Найбільшою проблемою методів, заснованих на словниках і правилах, є важкість процесу складання словника. Для одержання методу, що класифікує документ з високою точністю, терміни словника повинні мати вірну вагу, адекватну предметній області документа. Наприклад, слово «непередбачуваний» по відношенню до сюжету фільму є позитивною характеристикою, але негативною по відношенню до, наприклад, політика. Тому даний метод вимагає значних затрат часу людини, через те, що для хорошої роботи системи необхідно скласти велику кількість правил. Часто можлива автоматизація складення словників, проте зазвичай лише у дуже вузькій предметній області.

У простому вигляді тональний словник представляє з себе список слів зі значенням тональності для кожного слова. Приклад з бази ANEW, перекладений на українську мову, наведений в таблиці 2.1.

Таблиця 2.1 – Приклад тонального словника

Слово	Тональність (1–9)
happy	8.11
good	7.37
dull	2.85
angry	2.75
sad	1.51

2.3.4 Напівавтоматичне навчання

Напівавтоматичне навчання (semi-supervised learning) - це клас методів навчання, які використовують для тренувань як розмічені так нерозмічені дані, при чому як правило, кількість розмічених даних менше за кількість нерозмічених даних. Напівавтоматичне навчання залежить від навчання без учителя (без будь-яких помічених навчальних даних) та навчання з учителем (з повністю наголошеними навчальними даними). Багато дослідників, які навчаються в машинобудуванні, встановили, що немітовані дані, коли вони використовуються разом із невеликою кількістю мічених даних, можуть значно покращити точність навчання. Для отримання розмічених даних для вивчення проблеми часто потрібен кваліфікований агент-людина (наприклад, для транскрибування аудіо сегмента) або фізичний експеримент (наприклад, визначення тривимірної структури білка або визначення того, чи існує масло в певному місці). Таким чином, витрати, пов'язані з процесом маркування, можуть стати занадто високими, а отримання немаркованих даних є

відносно недорогим. У подібних ситуаціях напівавтоматичне навчання може мати велике практичне значення.

В аналізі тональності напівавтоматичне навчання не дуже розповсюджене [8], найбільш відомим методом напівавтоматичного навчання в аналізі тональності є метод, описаний у роботі Хе Юлан та Жоу Деу. [12] Метод полягає у використанні як словникових методів, так і класичного класифікатора, наприклад SVM. Словникові методи використовуються для навчання класифікатора на великій нерозміченій вибірці, і таким чином поступово класифікатор навчається особливостям предметної області.

2.4 Вибір показники ефективності роботи алгоритмів

В якості метрики правильності класифікації текстів були обрані точність (precision), повнота (recall), доля правильних відповідей (accuracy). Точність в межах класу – це частка текстів, що дійсно належать даному класу, щодо всіх текстів, зарахованих класифікатором до цього класу. Повнота системи – відношення числа знайдених класифікатором текстів, що належать класу, до числа всіх текстів цього класу в тестовій колекції. В результаті класифікації текстів рецензій тестової вибірки, до класу позитивних рецензій правильно віднесені TP текстів, неправильно – FP, до класу негативних рецензій правильно були віднесені TN текстів, неправильно – FN. Іншими словами:

- TP – істинно-позитивне рішення;
- TN – істинно-негативне рішення;
- FP – хибно-позитивне рішення;
- FN – хибно-негативне рішення.

Тоді, відносно класу позитивних рецензій, точність *Precision* та повнота *Recall* визначаються за наступними формулами:

$$Precision = \frac{TP}{TP + FP} \tag{2.15}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.16}$$

$$Accuracy = \left(\frac{TP + TN}{P + N}\right) \tag{2.17}$$

2.5 Порівняння ефективності існуючих методів

Основною метою даного розділу є виявлення того, який саме з наявних методів для вирішення проблеми аналізу тональності дає найефективніший результат. Для оцінки ефективності роботи методу, використовуються метрики ефективності, що були наведені в попередньому підпункті.

Загалом, як було показано раніше, методи аналізу тональності поділяються на такі групи – методи навчання з учителем, semi-supervised методи та методи навчання без учителя (словникові).

Серед методів машинного навчання з учителем найчастіше виділяють методи наївного Баєса, максимальної ентропії та метод опорних векторів. [8] Оберемо найбільш популярні та ефективні з них – НБК та SVM. До мінусів поданих методів, як і загалом методів навчання з учителем, можна віднести те, що для реалізації методів машинного навчання з учителем необхідно мати два достатньо великих набори даних: перший набір даних називається навчаючою вибіркою, а другий набір даних – тестовою вибіркою. Часто предметна область не є розміченою, а отже дані методи вимагатимуть додаткових, при тому досить великих, зусиль для створення виборки. Серед плюсів – подані методи зазвичай мають більшу точність, та лише незначно менш швидкі за словникві (що досить важливо для обробки великих об’ємів

даних). [8] Можливе навчання на вибірці з іншої предметної області, однак якість роботи буде гіршою, у порівнянні зі звичайним способом навчання.

Варто зазначити, що у даній роботі для навчання методів НБК та SVM використовується вибірка IMDB [9], а для тестування цих, та інших методів використовується самостійно розмічена вибірка з 2000 коментарів з сайту новин HuffingtonPost.

Важливим кроком для методів машинного навчання з учителем є формування вектору ознак, тобто формату, в якому документ буде представлений для класифікатора. Як правило, в якості таких ознак використовуються наявність певних термінів та їх частота, що включає в себе уніграми, біграми чи їх комбінацію, інформація щодо частин мови, тобто кожному слову в реченні привласнюється певний тег частини мови (POS-теги), або заперечення. Як було показано іншими авторами, провели порівняння ефективності різних методів векторного представлення текстів для НБК та SVM, максимальну ефективність має представлення ‘feature presence’, тобто наявність певних термінів, а не їх частота чи більш складні формули. [6]

При використанні методу опорних векторів важливою задачею є вибір ядра методу. В якості ядра можуть бути обрані як поліноміальне (однорідне та неоднорідне), радіальна базисна функція або радіальна базисна функція Гаусса, або сигмоїд. Однак, не існує загального підходу до вибору ядра, а також параметрів цього ядра, тому у якості ядра було обране лінійне, як у більшості робіт, що вивчають методи аналізу тональності. [8]

Загалом, НБК та SVM за великого розміру навчальної вибірки мають майже однакову ефективність, при чому SVM швидше досягає піку своєї ефективності, однак потім трохи втрачає у точності через свою чутливість до викидів. [10]

Методи, основані на словниках для класифікації текстів використовують тональний словник. Для його створення, може використовуватися один з трьох методів: складання вручну, методи на основі корпусів та методи на основі словників. Мінусом цих методів у контексті задачі що вирішується є надзвичайна важкість

складання повного словника, адже коментарі часто містять неологізми та помилки у словах, які усі потрібно буде врахувати при створенні власного словника. У даній роботі використовується словник, складений Університетом Ілінойсу в Чикаго. [11] Очевидно, що подібний підхід бути мати невелику точність, через те, що він ігнорує не тільки зв'язок між словами, неологізми та помилки, але і можливість того, що у поданій предметній області слова будуть мати іншу тональність.

З semi-supervised підходів оберемо класичний, однак найбільш застосовуваний – підхід з двох етапів, перший з яких – класифікувати текст за допомогою словникового підходу, а потім використовувати класифіковані тексти для того, щоб навчити інший класифікатор. [12] У якості класифікатора оберемо НБК з використанням біграм. Таким чином класифікатор навчається не тільки класифікувати тексти за словами зі словника, а і словами з предметної області текстів. Цей підхід добре працює, якщо немає достатньої вибірки текстів з предметної області, або у лексика текстів часто змінюється, у ній наявні помилки в словах або неологізми. [12] Таким чином, навіть за наявності достатньої навчальної вибірки з предметної області, через деякий час вона може почати втрачати свою актуальність, а отже класифікатори, навчені на ній – свою ефективність.

Нижче наведений підсумок порівняння методів аналізу тональності (таблиця 2.2)

Таблиця 2.2 – Ефективність роботи існуючих алгоритмів

	Precision	Recall	Accuracy
Словниковий метод	0.59	0.685	0.681
НБК	0.643	0.771	0.772
SVM	0.65	0.769	0.77
НБК з використанням біграм	0.683	0.782	0.791
SVM з використанням біграм	0.684	0.778	0.793
Метод Хе Юлан та Жоу Деу (з класифікатором НБК)	0.653	0.78	0.781

Як можна побачити з табл. 2.2, найкращі результати показують методи машинного навчання з учителем, зокрема метод опорних векторів та метод наївного Баєса. Як і припускалося раніше, ці методи досить порівнювані за якістю на великій навчальній вибірці. Окремо виділимо semi-supervised метод Хе Юлан та Жоу Деу, адже він показав досить непогані результати, не зважаючи на те, що первинна класифікація виконувалася за допомогою словникового методу.

Приклади текстів, на яких класифікація одним методом була вдалою, а іншим ні наведені нижче на рисунках 2.7-2.8.

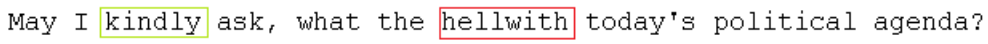

Рисунок 2.7 – Приклад неправильно класифікований словниковим методом, проте вдало – Наївним Байесом

Зеленим кольором виділено слово, що скоріше за все призвело до позитивної класифікації словниковим методом. Червоним – слово, завдяки яком НБК класифікував текст як негативний.

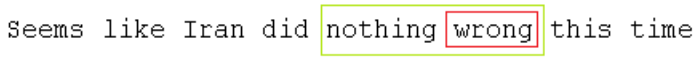

Рисунок 2.8 – Приклад неправильно класифікований НБК, проте вдало –НБК з використанням біграм

На рисунку 2.8 можна побачити, що два слова, які зазвичай використовуються в негативному контексті утворюють пару, що загалом має позитивний сенс. Цей приклад правильно розпізнається НБК з використанням біграм завдяки тому що його конструктивні особливості дозволяють не втрачати інформацію про слова з відстанню 1.

Треба також зауважити, що не зважаючи на те, що НБК та SVM мають приблизно однакові показники ефективності, останній навчається набагато повільніше.

Висновки за розділом

В даному розділі розглянуто, які основні методи використовуються для вирішення задачі аналізу тональності та показано, які основні недоліки присутні кожному методу. Окрім цього, в даному розділі проілюстровано, який основний підхід використовується для вирішення поставленої задачі в незалежності від обраних методів.

Також, обрано критерії, за якими буде в подальшому оцінюватись ефективність модифікованого методу. В якості таких метрик обрано точність та повноту, адже саме вони відображають правильність вирішення задачі текстової класифікації.

Продемонстровано існуючі варіанти представлення тексту, які впливають на результат вирішення задачі аналізу тональності. Визначено, які додаткові алгоритми можуть застосовуватись безпосередньо перед застосуванням основних алгоритмів.

Обгрунтовано вибір досліджуваних алгоритмів. Проведено порівняння їх ефективності за запропонованими критеріями. Наведено приклади, що наочно показують різницю у ефективності між алгоритмами.

3 АНАЛІЗ РЕЗУЛЬТАТІВ РОБОТИ

3.1 Аналіз запропонованого класифікатора

Виходячи з результатів попереднього розділу, у якості покращуваного методу оберемо semi- supervised метод з класифікатором НБК з біграмами. Зважаючи на те, що словникові методи, що використовуються на першому етапі мають не дуже високу точність, пропонується змінити перший етап та використовувати один з раніше обраних класифікаторів (НБК, SVM). Водночас, зважаючи на те, що точність цих двох методів за умови великої виборки порівнювана, якщо не однакова, а швидкість навчання НБК набагато вища, пропонується використовувати НБК.

Таблиця 3.1 – Ефективність роботи запропонованих модифікацій методу Хе Юлан та Жоу Деу за запропонованими критеріями

Модифікація методу	Precision	Recall	Accuracy
НБК + НБК	0.654	0.78	0.781
НБК + НБК з біграмами	0.661	0.778	0.783
НБК з біграмами + НБК з біграмами	0.683	0.789	0.799

Можна побачити (таблиці 2.2, 2.3), якщо запропонований метод на першому та другому рівні використовуватиме НБК, то його точність незначно покращиться відносно класичних методів НБК та SVM, застосованих на іншій навчальній вибірці.

У випадку коли лише другий рівень буде використовувати біграми, його показники ефективності зміняться лише незначно, можна припустити, що класифікатор на верхньому шарі не взмозі вивчити закономірності зв'язків між

сусідніми словами. Останній варіант – застосування у обох шарах методів НБК з використанням біграм має суттєво більшу ефективність за конкурентів.

Нижче наведений приклад тексту, де використання стандартних методів мало гірші результати за модифікацію, запропоновану у даній роботі (рисунок 3.1).

Apparently now AT&T is admitting it paid money to Cohen's Essential Consultants, the company formed to brbe women. This of course was happening while they were asking for approval of a merger with Time Warner.

Рисунок 3.1 - Текст неправильно класифікований НБК з біграмами, проте вдало – модифікованим методом

Можна припустити, що слово ‘bribe’ рідше зустрічається у відгуках на фільми, серіали чи ігри, але досить часто у коментарях до новин, тому, відповідно, саме такий варіант його написання і не зустрівся у вибірці, за якою навчались класифікатори.

Але, не зважаючи на покращення у точності обробки треба також згадати про негативну сторону подібних методів – сильне зменшення швидкодії відносно стандартних методів навчання з учителем (таблиця 3.2).

Таблиця 3.2 –Час роботи модифікацій методу Хе Юлан та Жоу Деу за умови 1 млн прикладів у навчальній вибірці як першого так і другого класифікаторів

Метод	Швидкість навчання (с на 1млн прикладів)	Швидкість класифікації (с на 1 млн прикладів)
НБК з біграмами	1159	1084
Модифікація (НБК + НБК)	1796	794
Модифікація (НБК + НБК з біграмами)	2060	1073

Продовження таблиці 3.2

Метод	Швидкість навчання (с на 1млн прикладів)	Швидкість класифікації (с на 1 млн прикладів)
Модифікація (НБК з біграмами + НБК з біграмами)	2303	1091

Наприклад, найкраща модифікація, у якій в обох шарах використовується НБК з використанням біграм працює майже (з точністю до похибки) у два рази повільніший у навчанні за звичайний НБК з використанням біграм, проте його ефективність роботи вища лише на 0.8% у Accuracy, 0.7% у Recall і 0% у Precision. У таблиці 3.3 нижче можна побачити порівняння найкращої модифікації по Accuracy, Recall, Precision з існуючими алгоритмами.

Таблиця 3.3 – Порівняння ефективності роботи існуючих алгоритмів і запропонованої модифікації

Метод	Precision	Recall	Accuracy
Словниковий метод	0.59	0.685	0.681
НБК	0.643	0.771	0.772
SVM	0.65	0.769	0.77
НБК з використанням біграм	0.683	0.782	0.791
SVM з використанням біграм	0.684	0.778	0.793
Метод Хе Юлан та Жоу Деу (з класифікатором НБК)	0.653	0.78	0.781
Запропонована модифікація	0.683	0.789	0.799

3.2 Обґрунтування вибору платформи та мови програмування

В якості мов та інструментів були обрані C#, python, nltk, scikit-learn.

Для додатка, що розроблюється був обраний вигляд Телеграм бота, що виконується у хмарі, тож важливим було обрати мову, що буде мати розроблену екосистему, у тому числі бібліотеку для простої взаємодії з API Телеграма, можливості легкої інтеграції з хмарою (Azure), зручні інструменти написання (Visual Studio) та буде достатньо популярною у розробці веб-застосунків, тому у якості основної мови був обраний C#.

C# - строго типізована об’єктно-орієнована мова з сильною типізацією, є частиною платформи .NET. Мова розроблена компанією Microsoft. Синтаксис C# близький до C++ і Java. Мова реалізує багато інструментів сучасного програмування - підтримує поліморфізм, перевантаження операторів, вказівники на функції-члени класів, атрибути, події, властивості, винятки, лямбда-вирази, коментарі у форматі XML.

В даний час C# використовується сотнями тисяч розробників. Згідно з рейтингом корпорації TIOBE, що базується на даних пошукових систем, у березні 2018 року C# знаходився на 5 місці серед мов програмування [13].

Суттєвим чинником при виборі мови python у якості основної мови для обробки даних та машинного навчання була наявність суттєвих наробків у сфері машинного навчання саме на цій мові, наприклад програмні бібліотеки NLTK, scikit-learn, що були використані нами у даній магістерській дисертації, а також бібліотеки Orange, розробленої Люблянським університетом.

Також, python був обраний через свою популярність і відносно високу швидкодію.

Бібліотека NLTK - це пакет бібліотек і програм для символьного і статистичної обробки природної мови. Містить графічні уявлення і приклади даних. Супроводжується великої документацією, включаючи книгу з поясненням основних

концепцій, що стоять за тими завданнями обробки природної мови, які можна виконувати за допомогою даного пакету. Бібліотека добре підходить для студентів, які вивчають комп'ютерну лінгвістику або близькі предмети, наприклад когнітивістику, штучний інтелект, інформаційний пошук і машинне навчання. NTLK часто використовується як навчальний посібник, як інструмент індивідуального навчання і в якості платформи для прототипування і створення науково-дослідних систем. [14]

scikit-learn - бібліотека, у якій реалізовані багато методів машиного навчання. Він має реалізації різних класифікаційних, регресійних алгоритмів і алгоритмів кластеризації, включаючи SVM, random forest, gradient boosting, k-means і таке інше. [15]

Python — це інтерпретована об'єктно-орієнтована мова програмування високого рівня з динамічною семантикою, що була розроблена в 1990 році Гвідо ван Россумом. Реалізовані структури даних високого рівня та динамічна семантика, динамічне зв’язування, автоматичне управління пам’яттю збільшують її зручність для швидкої розробки програм, а також для поєднання існуючих компонентів. Python підтримує модулі та пакети модулів, що сприяє повторному використанню коду.

3.3 Аналіз вимог користувача до програмного продукту

Надзвичайно велика кількість інформації у сучасному світі, у тому числі величезна кількість новин, що з’являється на різних новинних ресурсах, змушує людей виокремлювати новини та ресурси, які варті їх уваги. [16] Через високий темп життя, більша частина людей читає новини зранку та ввечері. [17] Отож, можна припустити, що досить велика частка з цих людей не хоче прокидатися, читаючи негативні (з їхньої точки зору) новини. Тому, приймаючи до уваги, що людина зазвичай читає ресурси, чиє бачення світу співпадає з їх власним, а отже можна

припустити , що і емоційне забарвлення коментарів під новинами буде співпадати з реакцією користувача.

Виходячи з вищезазначеного, подану проблему можна вирішити шляхом створення програмного забезпечення, що буде фільтрувати новини з обраних користувачем джерел, та, бажано, надсилати посилання на них у обраний час. Враховуючи популярність чат-ботів на час написання цієї наукової роботи, Телеграм-бот був обраний як оптимальне рішення.

3.4 Аналіз архітектури програмного продукту

Програмний продукт складається з чотирьох основних внутрішніх частин (рисунок 3.1):

- База даних, у якій зберігається результат роботи кожної з інших частин
- Parser – частина, що відповідає за обробку веб-сторінок новинних ресурсів та збереження статей новин, включно з назвою, текстом та коментарями у базі даних
- Telegram bot – частина, що відповідає за взаємодію з Telegram API – створення підписки на новини чи выдписку, а також саму розсилку
- Classifiers – частина, що відповідає за класифікацію коментарів на позитивні/негативні

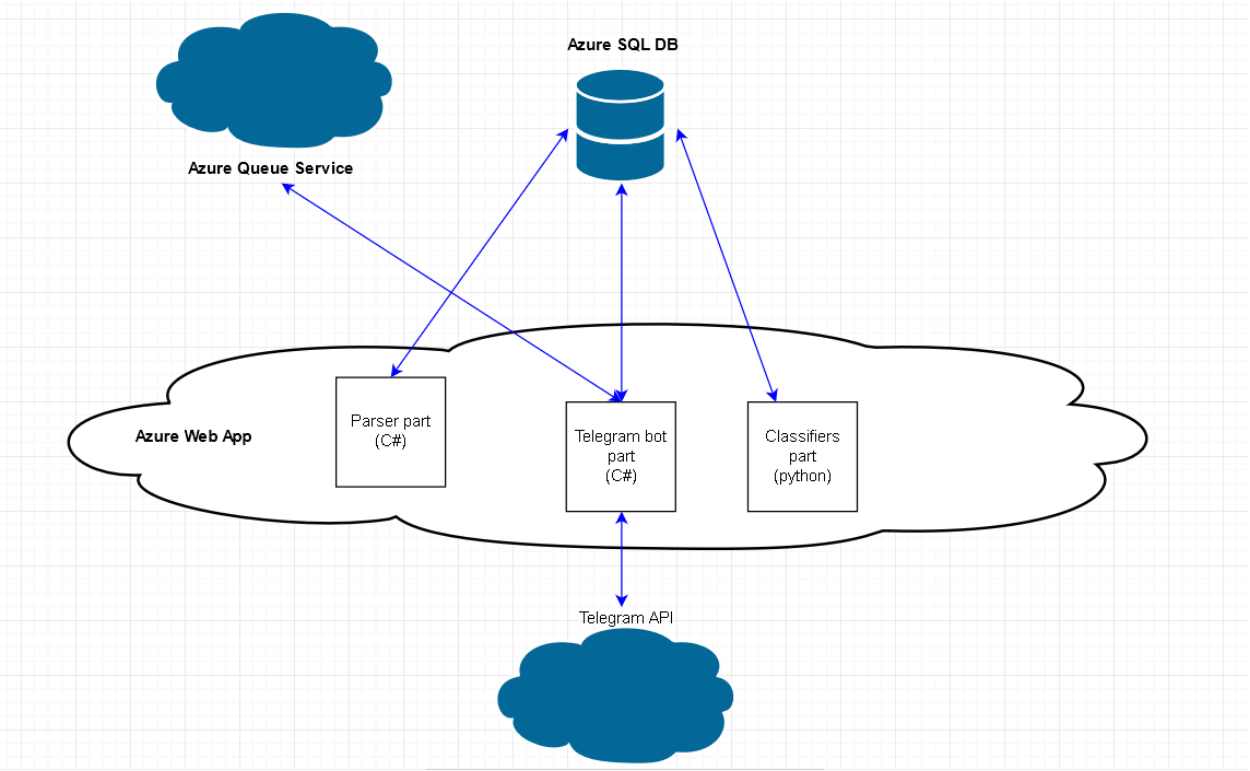

Рисунок 3.2 –Загальна архітектура системи

Усі чотири частини хостятся у Azure. Parser, Telegram Bot, Classifiers використовують Azure Web App – засобі для хостингу застосунків від Azure. Також, Telegram bot використовує Azure Queue Service для надсилання новин користувачам саме у бажаний час.

Також, варто уточнити, що хоча Telegram bot та Parser це окремі компоненти, проте вони мають значну частину спільного коду, що видповідає за обмін інформацією з іншими компонентами (наприклад базою даних).

Висновки за розділом

У третьому розділі основа увагу була зосереджена на аналізі розробленого класифікатора та архітектури, програмного продукту створеного на її основі.

Запропоновано декілька варіантів модифікації обраних алгоритмів. Проведено їх порівняння за запропонованими критеріями. Наведено приклад, що підтверджує ефективність його роботи.

Приводиться обґрунтування платформи та мови програмування (c# та Microsoft SQL Server для telegram та parser частини, python для реалізаціїкласифікаторів).

Проводиться аналіз вимог користувача до програмного продукту (способи використання, необхідні функції).

Аналізується архітектура програмного продукту. Описуються функції кожної частини програмного продукту.

4 РОЗРОБЛЕННЯ СТАРТАП-ПРОЕКТУ

4.1 Опис ідеї проекту

Основна ідея стартап проекту описана нижче у таблицях 4.1 – 4.22.

Таблиця 4.2 – Опис ідеї стартап-проекту

Зміст ідеї	Напрямки застосування	Вигоди для користувача
Телеграм-бот що агрегує новини з тем, що цікаві користувачам і надсилає обрану кількість новин, на які найкраще реагували у коментарях.	Бот може надсилати саме позитивні новини, з тем, що цікаві користувачам.	Дозволяє зменшити вплив поганих новин на настрій людей, що звикли просинатись, читаючи новини. Новини більш актуальні за більшість сайтів з «гарними новинами» і використовують обрані користувачем сайти, як джерела інформації.

Таблиця 4.3 – Визначення сильних, слабких та нейтральних характеристик ідеї проекту

п/п	Техніко-економічні характеристики ідеї	товари/концепції конкурентів			W (слабка сторона)	N (нейтральна сторона)	S (сильна сторона)
		Запропонований проект	Good news FOX	BBC Happy news			
.	Можливість обирати тематику новин	+	+	+	-	+	-
.	Актуальність новин	+	+/-	+/-	-	-	+
.	Можливість обрати декілька джерел інформації	+	-	-	-	-	+
.	Ціна	+/-	+/-	+/-	-	+	-

4.2 Технологічний аудит ідеї проекту

Таблиця 4.3 – Технологічна здійсненність ідеї проекту

п/п	Ідея проекту	Технології її реалізації	Наявність технологій	Доступність технологій
		Гібридна система класифікації	Розробити	Розроблено автором проекту
		Класифікаційні системи CRM, Naïve bayes	Наявна	Доступна
		Телеграм бот	Наявна	Доступна
Обрана технологія реалізації ідеї проекту: Для реалізації проекту було обрано CRM, Naïve bayes classifier, методи semi-supervised learning як основу гібридної системи класифікації коментарів до новин та телеграм бот для комунікації з користувачами.				

4.3 Аналіз ринкових можливостей запуску стартап-проекту

Таблиця 4.4 – Попередня характеристика потенційного ринку стартап-проекту

п/п	Показники стану ринку (найменування)	Характеристика
	Кількість головних гравців, од	Велика, немає точної кількості
	Загальний обсяг продаж, грн/ум.од	36 млрд $ (Загалом новини – 1.5 трлн $)
	Динаміка ринку (якісна оцінка)	Зростає
	Наявність обмежень для входу (вказати характер обмежень)	Недовіра до незнайомих джерел новин, відносна консервативність користувачів
	Специфічні вимоги до стандартизації та сертифікації	-

Продовження таблиці 4.4

п/п	Показники стану ринку (найменування)	Характеристика
	Середня норма рентабельності в галузі (або по ринку), %	10-15%

Таблиця 4.5 – Характеристика потенційних клієнтів стартап-проекту

п/п	Потреба, що формує ринок	Цільова аудиторія (цільові сегменти ринку)	Відмінності у поведінці різних потенційних цільових груп клієнтів	Вимоги споживачів до товару
	Бажання читати гарні новини зранку	Люди, що читають новини зранку	– Частота читання новин – Вік, консервативність у звичках – Бажання використовувати ботів	– Актуальність новин – Можливість обирати теми новин – Зручність використання

Таблиця 4.6 – Фактори загроз

№ п/п	Фактор	Зміст загрози	Можлива реакція компанії
1	Перехід людей на власне сайт новин	Замість використання боту люди будуть зразу переходити на один новинний сайт	Більш таргетоване обирання новин, збільшення кількості джерел для урізноманітнення новин
2	Збільшення ціни на хостинг в Azure	Збільшення ціни на хостинг в Azure	Перехід на власне апартне устаткування або у інше хмарне середовище

Таблиця 4.7 – Фактори можливостей

№ п/п	Фактор	Зміст можливості	Можлива реакція компанії
1	Зростання зацікавленості людей у новинах	Зростання купівельної спроможності людей і збільшення об'єму ринку	Збільшення реклами для збільшення долі ринку, окремий акцент на молоду аудиторію.
2	Зростання покоління «меленіалів»	За дослідженнями «меленіали» не хочуть витрачати час на нецікаві їм теми, частіше уникають «поганих» новин	

Таблиця 4.8 – Ступеневий аналіз конкуренції на ринку

Особливості конкурентного середовища	В чому проявляється дана характеристика	Вплив на діяльність підприємства (можливі дії компанії, щоб бути конкурентоспроможною)
1. Вказати тип конкуренції – монополія / олігополія / монополістична / чиста	Конкурентний ринок	Потреба у постійному розширенні функціоналу, покращенню UX, аналізі ринку і даних
2. За рівнем конкурентної боротьби – локальний / національний /…	Глобальний	
3. За галузевою ознакою – міжгалузева / внутрішньогалузева	Внутрішньогалузева	
4. Конкуренція за видами товарів: – товарно-родова – товарно-видова – між бажаннями	Товарно-видова	

Продовження таблиці 4.8

Особливості конкурентного середовища	В чому проявляється дана характеристика	Вплив на діяльність підприємства (можливі дії компанії, щоб бути конкурентоспроможною)
5.За характером конкурентних переваг - цінова / нецінова	Нецінова	
6. За інтенсивністю - марочна/не марочна	Марочна	

Таблиця 4.9 – Аналіз конкуренції в галузі за М. Портером

Складові аналізу	Прямі конкуренти в галузі	Потенційні конкуренти	Постачальники	Клієнти	Товари-замінники
	BBC Happy news, FOX Good news,…	-	-	Дуже багато	Ні
Висновки:	Є великими конкурентами, але «гані новини» часто специфічні або старі, мало оновлюються	-	-	Рівень чутливості до зміни цін, зручності інтерфейлу і якості UX	-

Таблиця 4.10 – Обґрунтування факторів конкурентоспроможності

№ п/п	Фактор конкурентоспроможності	Обґрунтування (наведення чинників, що роблять фактор для порівняння конкурентних проектів значущим)
1	Швидка агрегація новин	Бажання користувачів читати останні новини
2	Можливість обирати тематику новин	Цікавість користувачів саме до певної тематики
3	Маленька компанія	Швидка зміна стратегії

Таблиця 4.11 – Порівняльний аналіз сильних та слабких сторін «назва проекту»

№ п/п	Фактор конкурентоспроможності	Бали 1-20	Рейтинг товарів-конкурентів у порівнянні з поданим продуктом						
			–3	*–2*	*–1*	*0*	*+1*	*+2*	*+3*
1	Актуальність новин	15		X					
2	Можливість обирати теми/джерела новин	20	X						
3	Зручність використання	10				X			
4	Популярність компанії	1							X

Таблиця 4.14 – SWOT- аналіз стартап-проекту

Сильні сторони: актуальність новин, можливість обирати тематику новин	Слабкі сторони: популярність компанії
Можливості: зростання зацікавленості людей у новинах, зростання покоління «меленіалів»	Загрози: перехід людей на використання власне новинного сайту, збільшення ціни на хостинг в Azure

Таблиця 4.13 – Альтернативи ринкового впровадження стартап-проекту

№ п/п	Альтернатива (орієнтовний комплекс заходів) ринкової поведінки	Ймовірність отримання ресурсів	Строки реалізації
1	Проведення рекламних компаній	Середня	1-3 місяця
2	Розширення стандартного функціоналу (інтелектуальний підбір за змістом статей)	Середня	5 місяців

4.4 Розроблення ринкової стратегії проекту

Таблиця 4.14 – Вибір цільових груп потенційних споживачів

№ п/п	Опис профілю цільової групи потенційних клієнтів	Готовність споживачів сприйняти продукт	Орієнтовний попит в межах цільової групи (сегменту)	Інтенсивність конкуренції в сегменті	Простота входу у сегмент
1	Люди 16-18 років	Висока	Середній	Середня	Середня
2	Люди 18-25 років	Висока	Високий	Висока	Складна
3	Люди 35-45 років	Висока	Високий	Висока	Складна
4	Старше 45 років	Середня	Середній	Низька	Дуже складна
Які цільові групи обрано: 2, 3					

Таблиця 4.15 – Визначення базової стратегії розвитку

п/п	Обрана альтернатива розвитку проекту	Стратегія охоплення ринку	Ключові конкурентоспроможні позиції відповідно до обраної альтернативи	Базова стратегія розвитку*
		Визначити потреби кожної з цільових груп потенційних клієнтів, для кожної з них спеціально розробити стратегії приваблення клієнтів та маркетингової комунікації здійснювати продуктові новації	– Орієнтованість на кінцевого споживача – Швидкість «орієнтації на ринок»	Стратегія спеціалізації

Таблиця 4.16 – Визначення базової стратегії конкурентної поведінки

№ п/п	Чи є проект «першопрохідцем» на ринку?	Чи буде компанія шукати нових споживачів, або забирати існуючих у конкурентів?	Чи буде компанія копіювати основні характеристики товару конкурента, і які?	Стратегія конкурентної поведінки*
1	Ні	Забирати існуючих	Ні	Стратегія заняття конкурентної ніші

Таблиця 4.17 – Визначення стратегії позиціонування

п/п	Вимоги до товару цільової аудиторії	Базова стратегія розвитку	Ключові конкурентоспроможні позиції власного стартап-проекту	Вибір асоціацій, які мають сформувати комплексну позицію власного проекту (три ключових)
	- Актуальність новин - Можливість обирати тематику новин - Позитивне забарвлення новин	Стратегія заняття конкурентної ніші	По іміджу	– Позитивність – Актуальність – Кастомізація

4.5 Розроблення маркетингової програми стартап-проекту

Таблиця 4.18 – Визначення ключових переваг концепції потенційного товару

№ п/п	Потреба	Вигода, яку пропонує товар	Ключові переваги перед конкурентами (існуючі або такі, що потрібно створити
1	Актуальність новин	Новини, що пропонуються є більш актуальними за новини прямих конкурентів	- Більш свіжі позитивні новини
2	Можливість обирати тематику новин	Можливість обирати тематику позитивних новин, що цікава користувачеві	- Можливість обирати тематику позитивних новин, що цікава користувачеві

Таблиця 4.19 – Опис трьох рівнів моделі товару

<table>
<tr><td>Рівні товару</td><td colspan="3">Сутність та складові</td></tr>
<tr><td>I. Товар за задумом</td><td colspan="3">Бажання отримати товар не виходячи з дому, повторювана покупка товарів.</td></tr>
<tr><td rowspan="6">II. Товар у реальному виконанні</td><td>Властивості/характеристики</td><td>М/Нм</td><td>Вр/Тх /Тл/Е/Ор</td></tr>
<tr><td>1. Позитивність новин
2. Актуальність новин
3. «Влучність» новин відносно обраних тематик
4. Кількість реклами</td><td>Нм
Нм
Нм

М</td><td>Тл/Е
Тл/Е
Тл/Е

Вр/Е</td></tr>
<tr><td colspan="3">Якість: стандарти відсутні. Проект покрито тестами, регулярно тестується тестувальником.</td></tr>
<tr><td colspan="3">Пакування: Ні. Телеграм бот</td></tr>
<tr><td colspan="3">Марка: GN Bot</td></tr>
</table>

Продовження таблиці 4.19

Рівні товару	Сутність та складові
III. Товар із підкріпленням	До продажу: -
	Після продажу: -.
За рахунок чого потенційний товар буде захищено від копіювання: На ідею буде зареєстровано патент, прихильність користувачів до бренду.	

Таблиця 4.50 – Визначення меж встановлення ціни

№ п/п	Рівень цін на товари-замінники	Рівень цін на товари-аналоги	Рівень доходів цільової групи споживачів	Верхня та нижня межі встановлення ціни на товар/послугу
1	Безкоштовно (реклама)	-	Будь-який	-

Таблиця 4.61 – Формування системи збуту

№ п/п	Специфіка закупівельної поведінки цільових клієнтів	Функції збуту, які має виконувати постачальник товару	Глибина каналу збуту	Оптимальна система збуту
1	Цільові клієнти – переважно молоді люди, люди, що мають швидкий темп життя (зайняті люди), що звикли просинатись, читаючи новини.	Встановлення контактів із споживачами і підтримання їх. Формування попиту. Дослідницька робота зі збору маркетингової інформації.	Канал нульового рівня	Організація збуту всамотужки

Таблиця 4.72 – Концепція маркетингових комунікацій

№ п/п	Специфіка поведінки цільових клієнтів	Канали комунікацій, якими користуються цільові клієнти	Ключові позиції, обрані для позиціонування	Завдання рекламного повідомлення	Концепція рекламного звернення
	Цільові клієнти – переважно молоді люди люди, що мають швидкий темп життя, читають новини зранку і бажають не псувати собі цим настрій.	SMM Онлайн реклама: google, соціальні мережі	Інноваційність, зручність. По іміджу.	Створення відчуття зручності, лояльності до бренду. Завоювання нових клієнтів.	«New day is full of good things. We can tell»

Висновки за розділом

Основна мета даного стартап-проекту – дати користувачам можливість отримувати позитивні та актуальні новини зранку з джерел та тем, визначених самим користувачем.

Телеграм-бот що агрегує новини з тем, що цікаві користувачам і надсилає обрану кількість новин, на які найкраще реагували у коментарях.

Існує можливість ринкової комерціалізації за рахунок росту ринку та збільшення попиту на персоналізовані новини.

Конкуренція у сфері новин досить жорстка, проте є непогані шанси зайняти певний відсоток ринку за рахунок переваг поданого продукту. Поріг входження вцілому на ринок новин відносно високий, однак ніша персоналізованих новин ще розвивається, тому саме у цій сфері поріг знижений.

Подальша імплементація стартап-проекту є доцільною та рентабельною. Доцільним є також подальший розвиток розробленої технології класифікації, як однієї з найбільших конкуретних переваг на ринку.

ВИСНОВКИ

В даній роботі розв’язувалась задачі побудови системи для оцінювання емоційного відклику за коментарями. В роботі отримані наступні результати:

- Проведено аналіз існуючих методів (машинне навчання з учителем, без вчителя, заснований на словниках та правилах, напівавтоматичне навчання) та алгоритмів (НБК, SVM, алгоритм Хе Юлан та Жоу Деу) аналізу тональності. В результаті аналізу для даної роботи були обрані алгоритми НБК, SVM та алгоритм напівавтоматичного навчання, запропоновий Хе Юлан та Жоу Деу.
- Проведено аналіз сучасних літературних джерел та існуючих рішень. Аналіз показав, що більшість дослідників зосередили свою увагу на аналізі тональності повідомлень твітеру та оцінювання різних предметних виборок (оцінювання техніки, фільмів, ресторанів). Жодна з оглянутих систем не надавала можливості оцінювати емоційний відклик на новини за коментарями користувачів.
- Проаналізовано переваги та недоліки алгоритмів НБК, SVM та методу напівавтоматичного навчання, запропонованого Хе Юлан та Жоу Деу. Відповідно до предметної області було декількома способами модифіковано алгоритм Жоу Деу, проведено порівняння варіантів модифікації, обрано найкращий.
- Було запропоновано архітектуру програмного продутку, що включає модуль парсингу, модуль класифікації та модуль чат бота.
- На основі запропонованої архітектури було розроблено програмний продукт (телеграм бот), який надсилає новини з сайту обраного користувачем, у яких найбільший відсоток коментарів з позитивною тональністю.
- Обґрунтовано використання власного підходу до аналізу тональності

ПЕРЕЛІК ПОСИЛАНЬ

1. Opinion Mining on the Web by Extracting Subject-Aspect-Evaluation Relations [Електронний ресурс]: [Веб-сайт]. – Електронні дані. – Nozomi Kobayashi, Ryu Iida, Kentaro Inui, Yuji Matsumoto – 2006 – Режим доступу: https://www.aaai.org/Papers/Symposia/Spring/2006/SS-06-03/SS06-03-018.pdf
2. The Importance of Neutral Examples for Learning Sentiment [Електронний ресурс]: [Веб-сайт]. – Електронні дані. – Koppel, Moshe; Schler, Jonathan – 2005 – Режим доступу: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.84.9735&rep=rep1&type=pdf
3. Bing Liu. Sentiment Analysis and Subjectivity [Електронний ресурс]: [Веб-сайт]. – Електронні дані. – Bing Liu – 2010 – Режим доступу: https://www.cs.uic.edu/~liub/FBS/NLP-handbook-sentiment-analysis.pdf
4. Sentiment strength detection in short informal text / [Mike Thelwall, Kevan Buckley, Georgios Paltoglou et al]. // Journal of the American Society for Information Science and Technology.-2010.- No61.- P. 2544–2558
5. Making objective decisions from subjective data: detecting irony in customer reviews. [Електронний ресурс]: [Веб-сайт]. – Електронні дані. – Reyes Antonio, Rosso Paolo – 2012 – Режим доступу: https://www.researchgate.net/publication/257015931_Making_objective_decisions_from_subjective_data_Detecting_irony_in_customer_reviews
6. Feature Selection and Weighting in Sentiment Analysis [Електронний ресурс]: [Веб-сайт]. – Електронні дані. – O'Keefe, Tim; Koprinska, Irena – 2006 – Режим доступу: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.709.1463&rep=rep1&type=pdf

7. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews [Електронний ресурс]: [Веб-сайт]. – Електронні дані. – P. Turney. – 2002 – http://www.aclweb.org/anthology/P02-1053.pdf
8. Sentiment analysis algorithms and applications: A survey [Електронний ресурс]: [Веб-сайт]. – Електронні дані. – Medhata, Walaa; Hassan, Ahmed – 2014 – Режим доступу: https://www.sciencedirect.com/science/article/pii/S2090447914000550
9. Sentiment Analysis on Movie Reviews [Електронний ресурс]: [Веб-сайт]. – Електронні дані. – Режим доступу: https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews/data
10. A Comparison of the Accuracy of Support Vector Machine and Naıve Bayes Algorithms In Spam Classification [Електронний ресурс] : [Веб-сайт]. – Електронні дані. – Rita McCue, Jürgen Schmidhuber – 29.11.2009 – Режим доступу: https://pdfs.semanticscholar.org/55c5/9874114617b57eedd8636c5d0e7785fb885f.pdf
11. Opinion lexicon [Електронний ресурс]: [Веб-сайт]. – Електронні дані. – Режим доступу: http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar
12. Self-training from labeled features for sentiment analysis. Inf Process Manage [Електронний ресурс]: [Веб-сайт]. – Електронні дані. – He Yulan, Zhou Deyu – 2014 – Режим доступу: https://www.sciencedirect.com/science/article/abs/pii/S0306457310000932
13. TIOBE Index for March 2018 [Електронний ресурс]: [Веб-сайт]. – 2018 – Електронні дані. Режим доступу: https://www.tiobe.com/tiobe-index/
14. Text Classification using Naive Bayes [Електронний ресурс] – 2015. – Режим доступу: http://www.inf.ed.ac.uk/teaching/courses/inf2b/learnnotes/inf2b-learn-note07-2up.pdf
15. The importance of neutral examples for learning sentiment [Електронний ресурс] : [Веб-сайт]. – Електронні дані. – Rita McCue, Jonathan Schler – 21.10.2005 – http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.84.9735

16. Too much information [Електронний ресурс]: [Веб-сайт]. – Електронні дані. – Режим доступу: https://www.economist.com/node/18895468
17. How Americans get their news [Електронний ресурс]: [Веб-сайт]. – 2014 – Електронні дані. – Режим доступу: https://www.americanpressinstitute.org/publications/reports/survey-research/how-americans-get-news/